b/

Ack- Ho 5965/12 - KP 1373

THE JOURNAL OF PHILOSOPHY

UNIVERSITY LIBRARY
ALIPUR
CALCUTTA
19/12/0

12

VOLUME XCIX, NUMBER 1
JANUARY 2002

page
5 *Making Sense of the Senses: Individuating Modalities in Humans and Other Animals* Brian L. Keeley
29 *The Maxims Problem* Richard McCarty

BOOK REVIEWS
45 Lynne Rudder Baker: *Persons and Bodies: A Constitution View* Theodore Sider
49 John Rawls: *Lectures on the History of Moral Philosophy* Stephen Darwall

54 MEMO TO AUTHORS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721. You may also visit our website at: www.journalofphilosophy.org

K.P- 1373
dt: 05/8/03

THE JOURNAL OF PHILOSOPHY

2002

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 2002 by the Journal of Philosophy, Inc.

ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

No.

THE JOURNAL OF PHILOSOPHY

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J. E. Woodbridge and J. McKeen Cattell. In 1906, Wendell T. Bush became associated with the JOURNAL as co-editor. The title was changed in 1921, and the JOURNAL was incorporated in the State of New York under its present name in 1923. Since its founding, it has been published from the Department of Philosophy at Columbia University. It is typeset and printed by Cadmus Journal Services.

THE JOURNAL OF
PHILOSOPHY

VOLUME XCIX, NUMBER 1
JANUARY–DECEMBER 2002

Edited by
Brian Barry, Bernard Berofsky
Akeel Bilgrami, Arthur C. Danto
Kent Greenawalt, Patricia Kitcher
Philip Kitcher, Isaac Levi
Mary Mothersill, Philip Pettit
Carol Rovane, Achille C. Varzi

Consulting Editors
David Albert, John Collins
James T. Higginbotham, Charles D. Parsons
Wilfried Sieg

Managing Editor
Michael Kelly

Published by The Journal of Philosophy, Inc.

STATEMENT OF COPYRIGHT POLICY

We would like to draw your attention to the protection copyright affords against multiple photocopying by libraries, instructors, and copy centers. The 1976 copyright law provides that photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner—"fair use" meaning single copying for personal use. Accordingly, the JOURNAL authorizes personal or educational multiple copying only if a request is made in writing, and the fee of $.05 per page/per copy is paid directly to us. *Each copy must include a notice of copyright.* We shall turn over to the authors any substantial payments (over $30.00, an administrative fee) obtained from all photocopying. Inquiries about photocopying may be sent to the Managing Editor.

© Copyright 2002 by the Journal of Philosophy, Inc

ISSN 0022-362X

POSTMASTER Second class postage paid at New York, NY, and other mailing offices Please send address changes to THE JOURNAL OF PHILOSOPHY, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, NY, NY 10027 Fax (212) 932-3721

THE JOURNAL OF PHILOSOPHY

VOLUME XCIX, NO. 1, JANUARY 2002

19/12/10

MAKING SENSE OF THE SENSES: INDIVIDUATING MODALITIES IN HUMANS AND OTHER ANIMALS*

One of the perhaps most striking phenomenological facts about the human perceptual experience of the world is that it seems to be divided into *modes*. Our perceptions of the world are delivered to us in distinct classes, those of the separate sensory modalities, such as vision, hearing, and touch. When confronted with, say, a horse after a ride on a hot summer's day, we experience vivid impressions of the animal in the form of how it looks (large and brown), how it sounds (loud, heavy breathing), how it feels to the touch (hot and slick with sweat), and so on. That these different modes of experience are, well, *different* seems beyond question. The existence of separate sensory modalities would seem to be a brute fact about perception, if ever there were one.

The apparent "bruteness" of this fact of human experience might explain why, throughout history, philosophers and perceptual scientists have had relatively little to say about the nature of this difference. Yet the apparent fact of this difference lay at the very foundation of philosophical and scientific questions about the senses. Its very cen-

---

* I have received many useful suggestions on early drafts of this paper. In particular, I want to thank the following: Kathleen Akins, Eric Brown, Austen Clark, Carla Fehr, George Graham, Margaret Holland, Jim Lennox, Alva Noë, Peter Ross, Eric Saidel, David Sanford, and Virgil Whitmyer. Antecedents of this paper were presented at the 1999 meeting of the *Society for Philosophy and Psychology*, and at the 1999 Eastern Division meeting of the *American Philosophical Association*. At these meetings, I received thoughtful comments from my respondents, Tom Polger and Murat Aydede, respectively. Thanks go to audiences at the University of Northern Iowa, Iowa State, Duke University, Franklin and Marshall College, and the Center for the Ecological Study of Perception and Action at the University of Connecticut. During the writing of this article I was financially supported by the Philosophy/Neuroscience/Psychology Program at Washington University in St. Louis and the McDonnell Project in Philosophy and the Neurosciences.

0022-362X/02/9901/5–28

trality is what makes it simultaneously both a difficult and a crucial question: Exactly what is it that distinguishes the senses from one another?

I. INTRODUCTION

My goal here is two-fold. First, I want to bring attention to this underexplored topic—to bring together much of what little has already been written on it and, in doing so, to encourage others to take up the challenge it represents. Second, I want to propose an answer to the question. The answer is based on the criteria used in neuroethology, one of the sciences that daily confronts issues related to distinguishing, comparing, and contrasting sensory modalities. Distinguishing the senses in a way useful to those sciences which study perception requires knowledge of several factors, including the categories of physics, as well as the neurobiology, evolutionary biology, and behavior of the organism whose senses are to be distinguished. On my account, to possess a genuine sensory modality is to possess an appropriately wired-up sense organ that is historically dedicated to facilitating behavior with respect to an identifiable physical class of energy. Perhaps the correct way to think about modality is best suggested by a definition of the term in *Webster's Ninth New Collegiate Dictionary*: "One of the main avenues of sensation (as vision)." Modality is an "avenue into" an organism. Question: What travels on an avenue? Answer: information about the physical state of the world exterior to the *central nervous system* (CNS). What constitutes an "avenue"? An evolutionary dedicated sense organ that converts energy into nerve impulses and conveys those impulses to the CNS. This captures the original sense of the term: the different senses are different "modes" of perceptual interaction with the world.

I shall leave further discussion of my proposed answer to sections II (where I discuss how to set up the problem of distinguishing the senses) and III (where I present my proposal in detail). Perhaps surprisingly, I shall argue that the distinct experiential qualities of perception—the qualia—so central to the common-sense understanding of perception are simply nonstarters for a scientific understanding. I draw a similar conclusion about the most venerable theory of the senses: Aristotle's account of the senses in terms of the "proper objects of sensation." In section IV, I shall turn my attention to these prima facie plausible proposals in order to show how they fail to provide a firm foundation for an empirically adequate account of sensory differentiation.

Both the Aristotelian and the qualia-based approaches are best thought of as representing common-sense accounts of the senses.

Although they may well be adequate for this purpose, this is not my goal. To the contrary, I offer what might be termed a "thoroughly naturalized" approach to distinguishing the senses—one that proposes to make sense of the problem as it presents itself to the perceptual sciences. Therefore, one way of looking at what I do here is to provide an eliminative materialist theory of the sensory differentiation as an alternative to common-sense theories.

Before proceeding to the answers, however, I need to explain a little more fully what the *philosophical* question is here. What is it about the question of how we ought to differentiate the senses from one another which should draw the interest of philosophers? Further, what is the philosophical payoff?

Let us get the somewhat facetious answer out of the way first. If one defines philosophical problems as those questions which actual philosophers find important enough to grapple with, then the issue of distinguishing the senses would seem to qualify as sufficiently philosophical. Although, as I note above, the topic has received relatively little philosophical interest, it has nonetheless drawn the interest of philosophers as varied as Aristotle, H. P. Grice, David M. Armstrong, and Nelson Goodman. Introducing his take on the issue, John Heil[1] observes: "Regrettably, philosophers have had little to say about what distinguishes the senses from one another" (*ibid.*, p. 3). That situation may be changing, however, as evidenced by a number of recent papers.[2]

Surely, we can say more about the source of this interest. Turning again to Heil: "An explication of the senses, a determination of what constitutes a sensory mode must, it seems, occupy a position of central importance in any theory of perception" (*op. cit.*, p. 3). Perceptual psychology, sensory biology, neuroethology, and numerous other sciences make foundational use of the notion of separate senses without much discussion of the details. Such sciences typically just assume the existence of different senses. One sees this reflected in the textbooks used in undergraduate psychology and neuroscience curricula. These textbooks are often divided into chapters covering the various separate senses, but it is rare to see an explanation for what makes the separate senses separate. (Not that this is a feature unique to scientific theorizing. Philosophical theories of perception likewise tend to start from an unexplicated foundation of differentiated senses. Those which do explicate such a foundation typically do so in terms of the approaches I shall criticize in section IV.)

---

[1] *Perception and Cognition* (Berkeley: California UP, 1983).

[2] Norton Nelkin, "Categorising the Senses," *Mind and Language*, v (1990): 149-65; Dominic M. McIver Lopes, "What Is It Like to See with Your Ears? The Representational Theory of Mind," *Philosophy and Phenomenological Research*, LX (2000): 439-53; Peter W. Ross, "Qualia and the Senses," *Philosophical Quarterly*, LI (2001): 495-511; Alva Noë, "On What We See," *Pacific Philosophical Quarterly*, LXXXIII (in press).

I agree with Norton Nelkin: the question of differentiating the senses "asks how we should define the senses so as to make them scientifically useful concepts. More metaphysically, [it] asks what is the *real* nature of the senses" (*op. cit.*, p. 149). That is, the issue engages philosophy of psychology in both of its emphases. To the extent that philosophy of psychology is a branch of philosophy of science, the notion of the senses as differentiated from one another is a core notion in perceptual sciences; a scientific assumption requiring philosophical justification. Philosophy of psychology also has a foot in philosophy of mind, and Nelkin's "more metaphysical" aspect of the issue invokes that connection. Part of what is required in spelling out this fundamental aspect of perception involves explicating exactly what sort of metaphysical entity a sensory modality is.

In discussing this topic with colleagues, I have discovered a surprising amount of unrecognized disagreement among individuals. Some believe that the problem is intimately tied up with qualitative experience, to the extent that a "nonconscious sensory modality" is seen as an oxymoron. To have a sense is to have a unique set of experiences—qualia—associated with that sense. But others see the issue as much more related to the sorts of things in the world which can affect behavior, regardless of whether we are ever consciously aware of such influence. For example, they feel that we would be right to posit a human magnetic sense if people could be shown to respond behaviorally to magnetic stimuli in a systematic fashion regardless of whether we experience "magnetic qualia." Still others insist that the senses are strictly related to sense *organs*; for example, refusing to accept the existence of a genuine "vestibular sense" until informed of the existence of the semicircular canals and associated cranial nerves. "No organ, no sense" seems to be their rule.

These differing intuitions agree on Aristotle's five senses, so people tend not to notice their different intuitions. Such apparently widespread agreement reinforces the strength with which these intuitions are held. After echoing the sentiment that the differentiation of the senses have "seldom been directly discussed," J. W. Roxbee Cox[5] goes on to diagnosis this curious condition by supposing that it may be "due to the fact that certain answers have seemed to different people so obvious that they did not need to be supported by argument" (*ibid.*, p. 530). The disagreement among individuals begins to become clear once we start considering potential senses in addition to Aristotle's five, examples of which I shall be discussing below. What makes the problem of discriminating the senses philosophically interesting is that the intuitions here are *fundamental*, *firm*, and generally *unrecognized* by the people who hold them. When one finds a disagreement with these three qualities, it is a clarion call for philosophical inquiry.

---

[5] "Distinguishing the Senses," *Mind*, LXXIX (1970): 530-50.

That I shall argue that qualia are *not* central to differentiating the senses indicates one important philosophical payoff from the topic under discussion here. The legitimacy of the concept of qualia in our scientific understanding of the mind is a matter of ongoing controversy. One line of argument might run thus:

P1: Qualia are essential to the project of differentiating the senses; one cannot make sense of different modalities except by reference to the qualitative character of perceptual experience.
P2: Differentiating the senses is an essential component of any science of perception, such as perceptual psychology or sensory biology.

∴ C: Qualia are an essential component of any science of perception.

I fully accept P2. In offering my "qualia-free" account of sense differentiation, however, I will be attacking P1—an attack that, if successful, can only undermine one of the primary purposes for which qualia could be useful. But if the account I defend here is wrong, and qualia are essential to differentiating the senses, then this would represent a powerful, new argument for the scientific legitimacy, nay the scientific *necessity*, of qualia.

For these reasons, I suggest that it is clear that the problem of how to differentiate the sensory modalities is an interesting and important philosophical topic. I shall return to the topic of philosophical implications in the final section. That said, assuming I have adequately justified the legitimacy of the topic, I shall consider in the next section how to address the question and consider what form the answer should take.

II HOW TO THINK OF THE QUESTION AND THE FORM OF ITS ANSWER

Aristotle famously counted five human sensory modalities: vision, hearing, taste, smell, and touch. Since then, his list has been expanded to include such potentially novel modalities as proprioception (our sense of the relative location of our limbs) and the vestibular sense (the sense of one's orientation with respect to gravity). So, Aristotle undercounted, but by how many? Reviewing the human sensory literature, Robert Rivlin and Karen Gravelle[4] opine: "Five [is] obviously just not enough to account for the huge range of sensory possibilities of which the human species is capable; *seventeen* senses is probably a more accurate count" (*ibid.*, p. 17; my emphasis).

Lurking behind the empirical seeming question of the number of modalities, we find a philosophical question of *how we ought to count them*. By what criterion (or set of criteria) should we count individual senses? I believe it is wise to recognize two different versions of this problem. The first version I call *Aristotle's problem*: How many modalities do humans have and how ought we decide the issue? This version of the problem is rather hoary, and I do not believe I need to introduce it further.

The same problem arises again, in a slightly different guise, in animal sensory biology. Aristotle not only counted a total of five *human* senses, he thought that was the number found in the entire animal kingdom. We also have good reason to believe, however, that he underestimated the sensory skills of other species. The animal kingdom is full of all sorts of wonderful ways of detecting the world. Certain snakes and boas use special pits below their eyes to sense infrared.[5] It has been proposed that certain fish, such as shad, can hear in the ultrasonic frequency range,[6] as can bats.[7]

While all these examples are interesting, one might argue that they are not genuinely different from human sensory capacities. If our ears or eyes, say, were built just a little differently, then we too could hear in the ultrasonic range or see in the infrared. Nonetheless, there are animals with sensory capacities that are extremely alien with respect to human capacities. Sharks are apparently capable of sensing magnetic fields.[8] There are also animals, including several species of fish, capable of perceiving electric potential.[9]

---

[4] *Deciphering the Senses. The Expanding World of Human Perception* (New York: Simon and Schuster, 1984).

[5] Theodore H. Bullock and Raymond B. Cowles, "Physiology of an Infrared Receptor: The Facial Pit of the Pit Viper," *Science*, cxv (1952): 541-43; reprinted in T. H. Bullock, *How Do Brains Work? Papers of a Comparative Neurophysiologist* (Boston: Birkhäuser, 1993), pp. 138-39.

[6] D. A. Mann, Z. Lu, M. C. Hastings, and A. N. Popper, "Detection of Ultrasonic Tones and Simulated Dolphin Echolocation Clicks by a Teleost Fish, the American Shad (*Alosa sapidissima*)," *Journal of the Acoustical Society of America*, civ (1998): 562-68.

[7] Donald R. Griffin and Robert Galambos, "The Sensory Basis of Obstacle Avoidance by Flying Bats," *Journal of Experimental Zoology*, LXXXVI (1941): 481-506; reprinted in L. D. Houck and L. C. Drickamer, eds., *Foundations of Animal Behavior Classic Papers with Commentaries* (Chicago: University Press, 1996), pp. 553-78.

Consider, for a moment, the star-nosed mole. Perhaps the most striking thing about this animal is the structure that gives the animal its name: a fleshy, tendriled nose. Casual observation quickly reveals that its star-shaped nose is most likely a sense organ. The mole constantly jabs it against the ground when exploring the environment. For the sensory biologist, a question arises: Given that the nose is a sense organ, what modality is involved? Is the nose a tactile receptor, a chemoreceptor, or perhaps even an electroreceptor? These questions have been a topic of recent debate in biology. In 1993, scientists[10] at the Smithsonian Institution published experiments suggesting that the star-nosed mole uses its nose for electroreception. Others[11] dispute this claim, and the recent work of Kenneth C. Catania and Jon H. Kaas[12] argues that the nose's modality is most likely tactile in nature.

What I shall call the *star-nosed mole problem* is the general philosophical problem raised by this type of scientific controversy. On what philosophical grounds should we decide which organisms possess which modalities? When scientists claim to have discovered a new sensory modality, what is the theoretical content of this claim? If philosophers have paid relatively little attention to Aristotle's problem, they have almost completely ignored this related problem.

The above discussion gives some feel for what I take the *question* to be. But before continuing, I shall say a little more about my assumptions concerning the appropriate form of the *answer*. I introduced the topic by proposing that we should think of the sensory modalities as "avenues" into the organism for information about the physical state of the external world. Further, I draw the boundary at the edge of the CNS. At this point, three questions arise.

---

[8] A. J. Kalmijn, "Electric and Magnetic Field Detection in Elasmobranch Fishes," *Science*, CCXVIII (1982): 916-18.

[9] See my "Fixing Content and Function in Neurobiological Systems: The Neuroethology of Electroreception," *Biology and Philosophy*, XIV (1999): 395-430; and "Shocking Lessons from Electric Fish: The Theory and Practice of Multiple Realization," *Philosophy of Science*, LXVII (2000): 444-65.

[10] E. Gould, W. McShea, and T. Grand, "Function of the Star in the Star-nosed Mole, *Condylura cristata*," *Journal of Mammalogy*, LXXIV (1993): 108-16.

[11] P. A Schlegel and P B. Richard, "Behavioral Evidence against Possible Subaquatic Electrosensitivity in the Pyrenean Desman *Galemys pyrenaicus* (Talpidae, mammalia)," *Mammalia*, LVI (1992): 527-32.

[12] "The Organization of the Somatosensory Cortex of the Star-nosed Mole," *Journal of Comparative Neurology*, CCCLI (1995): 536-48; "The Unusual Nose and Brain of the Star-nosed Mole: A Star in the Brain," *BioScience*, XLVI (1996)· 578-86.

First, why draw a boundary at all? Central to the concept of perception is that there is a duality: a perceiver and a perceived. Senses do not exist *en vacua*; they are possessed by "sensers" and, at the most basic level of analysis, they act as some kind of informational connection between the world and the psychological entity that possesses them. A theory of perception must posit such a boundary between perceiver and perceived. The concept of modality would then refer to the different ways in which information about the world crosses that boundary.

Second, if we must draw such a boundary, is it not necessarily arbitrary? Why draw the boundary at one place versus another? My answer is that drawing the boundary at the CNS is not arbitrary if we keep in mind my initial starting point: I am here presenting a concept of sensory modality which is useful to sciences that deal with human and animal perception. Given an environment and some kind of psychological system under investigation—two things I am assuming here—a sensory modality is a potential mode by which information in the environment can pass through some boundary and enter into the psychological system. If we are materialists and believe that the central nervous system is somehow the locus of the mind, then it is not arbitrary to draw the boundary between the environment and the psychological system at the "world/CNS" junction.

Finally, I claim that the appropriate boundary surrounds the CNS and *not* the body in general. I do not have space to go into an example here, but I wish to leave open the possibility of "internal modalities"—for example, proprioception or our sense of thirst (mediated by hypothalamic blood osmotic pressure receptors). The point in these examples is that information about the world external to the CNS (in this case, the body) is passing across the conceptual boundary surrounding the CNS along a particular pathway that represents a mode of interaction different from other pathways.

III FOUR PROPOSED CRITERIA FOR INDIVIDUATING THE SENSES

So much for preamble. Now, I shall survey the four criteria that I believe, taken together, constitute the best account of modality differentiation. These criteria are, I propose, individually necessary and jointly sufficient. I shall discuss why each is a necessary component of a correct account, as well as what is insufficient about each taken on its own.

(1) *Physics*: the external physical conditions upon which the senses depend. That is, we might distinguish the senses by reference to the physical qualities of their respective stimuli: vision is the detection of differences in electromagnetic stimuli; olfaction is the detection of differences in concentrations of chemical stimuli.

Sensory systems operate by physiologically responding to different forms of energy in the environment. Independent of any psychological and biological concerns, physics provides us with an ontology of energy forms. According to our best physics, electromagnetic phenomena are just different from mechanical energy, both of which are different from chemical gradients. This ontology provides a foundation for a nonarbitrary sensory differentiation.

In providing an ontology of forms of energy, what physics provides is the space of *possible* modalities. Whether any animal on earth makes use of a magnetic modality or not, the fact that physics identifies magnetism as a type of energy raises the possibility of a magnetic modality. Although necessary, physics is not a sufficient criterion. By itself, physics tells us nothing about which modalities *actually* exist. Physics can give us magnetism as a form of energy, but in order for there to be a magnetic modality, there must be organisms capable of sensing this physical class of stimuli.

> (2) *Neurobiology*: the character of the putative sense organs and their modes of connection with the brain.[13] For example, vision is what we do with our eyes; audition is what we do with our cochlea and associated auditory brain areas.

This would seem to be the additional "contribution of the organism" required by the discussion of (1). It also matches well with some of our naive notions of modality individuation; that is, we individuate our modalities in part on the basis of our sense organs. As Armstrong[14] points out, however, individuating sense organs is itself no mean problem. Following a suggestion by Anthony Kenny, Armstrong proposes (*ibid.*, pp. 211-13) that sense organs are bodily structures that we actively use to gain information about the world, as when we open and move our eyes to see or cock our head to hear. But he continues, this runs up against the problem that we do not actively move organs in *all* the putative cases of sense. For example, we do not *do* anything to gain vestibular information. It seems to be ever present (which might explain why Aristotle did not remark upon it). The use of an organ in active perception does not seem to be of help here.

My suggestion is to follow sensory neurobiology and look for physiological, anatomical, and morphological characteristics to individuate sense organs. A legitimate sense organ, I suggest, has three characteristics:

---

[13] Some of the felicitous wording of criteria 2 and 3 are due to C. A. J. Coady's "The Senses of Martians," *The Philosophical Review*, LXXXIII (1974): 107-25, which is a commentary on Grice's "Some Remarks about the Senses," in R. J. Butler, ed., *Analytical Philosophy*, Series I (New York: Oxford, 1962), reprinted in Grice, *Studies in the Way of Words* (Cambridge: Harvard, 1989), pp. 248-68 (1989 printing cited here).

[14] *A Materialist Theory of Mind* (London: Routledge, 1968)

First, it has to be an organ that physiologically responds to a naturally occurring degree of physical stimulation.[15] If the organ in question is a genuinely magnetoreceptive organ, then it needs to respond to the presentation of appropriate amounts of magnetic stimulation.

Second, a sense organ needs to be innervated by neurons; in other words, something cannot be a sense organ unless it is "wired up" properly to the CNS. (Hence, the neurobiology criterion explicitly helps rule out vestigial sense organs.)

Third, we need to be able to identify an "end organ" of some type; the neurons leading to the CNS from the putative sense organ need to *start* there. This is to discount as sense organs the second, third, fourth, and so on neurons in the chain leading from the sensory periphery. A sense organ must contain cells that respond to energy in the environment, not another nerve cell. Most legitimate sense organs feature morphologically distinct end organs (for example, the rods and cones of the eye) that are physically so constituted as to transduce some class of energy (say, photons) into the electrochemical energy of neurons. But some senses seem to make use of so-called "free nerve endings" without any identifiable end *organ*. In these cases, however, the putative sensory cells still lie at the *beginning* of a chain of neurons leading to the CNS. In its usage here, the stress is on end rather than organ.

There is, of course, more to be said here, but suffice it to say, I think it is possible to deal with the issue of individuating sense organs. I said above that this criterion "helps" deal with the problem of vestigial organs. It does not completely rule them out, however, which indicates the insufficiency of this criterion. It is possible to conceive of an end organ which is wired up to the CNS, but which passes on information of which the animal never makes any use. To have a genuine modality, it is not enough to have an organ of a particular type wired up to the CNS; that organ has to allow you to do the right sorts of things. Hence, the next criterion:

> (3) *Behavior*: the ability to discriminate behaviorally between stimuli that differ only in terms of a particular physical energy type.

---

[15] It must to respond to a *naturally occurring* degree of physical stimulation. The reason for this additional lemma will be discussed under criterion (4), dedication, below.

Part of what it means to have a modality is to be able to make behavioral discriminations within that modality. Once again, the suggestion of this individuating criterion generates further individuation requirements. To wit, how ought we individuate "behaviors?" For example, it seems odd to identify tanning as a behavior. My plan here is simply to follow Fred Dretske[16] on this issue: "behavior is endogenously produced movement, movement that has its causal origins within the system whose parts are moving" (*ibid.*, p. 2). Tanning is something that *happens to* a person, not something one does. On the other hand, pressing a button or vocalizing are paradigmatic behaviors.

The science that is arguably most invested in the study of the relationship between behavior and the senses is the science of psychophysics. In his *Sensory Qualities*, Austen Clark[17] wields the impressive conceptual and empirical framework of psychophysics to ground an account of qualitative appearances which is materialist in spirit and which answers a variety of philosophical questions concerning the nature of appearances. Furthermore, he proposes that psychophysics alone has the resources to individuate the sensory modalities. For this reason, I shall now go into that proposal in some detail to show why the behavior criterion taken alone is insufficient to differentiate the senses.

Central to Clark's account is the psychophysical concept of "matching" (taken originally from Goodman[18]). "Matching" is the relation between two stimuli that differ physically but are nonetheless in principle indiscriminable from one another. For example, two color patches might reflect slightly different wavelengths of light, but differ so minutely that any human subject would report that the two patches are perceptually identical. Perhaps surprisingly, it turns out that the matching relation is nontransitive: stimulus *A* may match stimulus *B*, and stimulus *B* match stimulus *C*, but stimulus *A* need not match stimulus *C*. Using this relationship, one can construct "matching spans" of stimuli in which each stimulus matches its immediate neighbors, but the ends of the spans are easily discriminable.[19] For example, we can construct such matching spans for color, creating a series of stimuli that vary infinitesimally by wavelength from red to green, say. A given observer, when presented with any two adjacent elements from this series, will be unable in principle to discriminate them, even though she can clearly distinguish red from green (the ends of the matching span).

---

[16] *Explaining Behavior. Reasons in a World of Causes* (Cambridge: MIT, 1988).
[17] New York: Oxford, 1993.
[18] *The Structure of Appearance* (Boston: Reidel, 1977, 3rd edition).
[19] Clark, pp. 79-84.

This is only a tiny fraction of the story Clark tells, but it is all we need to understand his proposal for individuating modalities (also taken from Goodman):

> Facts about matching can individuate modalities. Sensations in a given modality are connected by the matching relation. From any sensation in the given modality, it is possible to reach any other by a sufficiently long series of matching steps. Distinct modalities are not so connected. One can get from red to green by a long series of intermediaries, each matching its neighbors; but no such route links red to C-sharp (*op. cit.*, pp. 140-41).

There are two problems with Clark's proposal. First, on this account of modality individuation, we get many more modalities than we might have otherwise thought. Clark's account would entail breaking up the modalities into many submodalities, for not only can you not get from red to C-sharp (thus demarcating vision from audition), you also cannot get from red to "moving left to right across the visual field" (or however motion sensations ought to be described), nor from C-sharp to "darn that's as loud as a 747 engine from ten feet away" (or however auditory intensity sensations ought to be described). Lacking the appropriate matching relationships, there is no reason to class the "color" and "motion" submodalities of vision as both being submodalities *of vision*. We are left with an account that makes "color" and "motion" as distinct from one another as each is distinct from "pitch." Such an account fails to provide the resources for grouping together as "the same modality" sensory qualities that we would intuitively group together. All Clark has shown us is how to individuate submodalities.

Second, Clark's account attributes the wrong modalities to the wrong organisms. For example, using (3) alone, humans would have an electrical modality! Consider the following: humans are easily capable of discriminating fully charged nine-volt batteries from "dead" ones, simply by sticking them to the tongue. Nine volts is more than enough electricity to stimulate the sensory cells of the tongue. You could do all sorts of interesting psychophysical studies of human electrical discrimination. Yet it seems absurd to claim that humans have an electrical modality, at least not in the same sense as electric fish are thought to. Something more is needed than simply a capacity to discriminate behaviorally stimuli of a certain physical type. We need some acknowledgement of the function of the alleged sensory modality in the species under consideration.

(4) *Dedication*: the evolutionary or developmental importance of the putative sense to an organism. For example, we ought not attribute an electrical modality to an individual unless electrical properties of the world are part of the normal environment of that individual and to which the organism is attuned.

Dedication, a concept taken from the science of neuroethology,[20] is an attempt to make relevant what is biologically important to an organism. Just because a particular individual can behaviorally respond to a particular class of stimuli does not give us warrant to propose a modality for sensing that class of stimuli. In the example above, the reason why it is absurd to attribute an electrical modality to humans is that, as a species, we do not go around using this electrical capacity of our tongues to sense the electrical properties of the world. Electric fish, on the other hand, detect the electrical properties of their world all of the time. It allows them to navigate the nearly opaque environment of the tropical waterways in which they live. It allows them to carry out a nocturnal lifestyle, which in turn gives them a fitness advantage over nonelectroreceptive fish.[21]

Neuroethologists, neurobiologists who study the evolution and neural basis of animal behavior, make a distinction that is useful here. The suffix *-detection* is applied to any organism that is capable of responding, by any means, to the presence of a particular type of stimulation in the environment. The suffix *-reception* is reserved for those organisms which carry out such sensory discriminations through the use of a dedicated anatomical system of structures. So, they would say that whereas electric fish are capable of electro*reception*—they can behaviorally respond to electrical stimulation using structures that have evolved specifically to process electrical information about the world—at best, humans are capable of electro*detection*—humans detect electrical potential using the pain, taste, and tactile receptors on the tongue. By considering both the developmental history of an individual and the evolutionary history of its species, we can determine to what forms of energy in the world a putative sense organ has become dedicated.

To consider a concrete example, there are at least three different things that can stimulate a vertebrate eye: (1) photons, (2) mechanical distortion (as when you press the eyeball with a finger), and (3) a properly inserted stimulating electrode (as in a neurophysiological experiment). These are quite clearly three different forms of energy (electromagnetic, mechanical, and electrical, respectively) to which the eye qua sense organ is physiologically responsive (satisfying criteria (1) and (2)). All three types of stimulation can elicit behavior as required by criterion (3). What makes the eye part of a visual system, but not part of a mechanosensory or stimulating-electrode-receptive system, is the evolutionary history of those vertebrates which have eyes. It is this history which determines to what sense a putative sense organ is dedicated. Dedication, in turn, allows us to distinguish those animals which genuinely possess a given sensory modality from those which have figured out a way of using some other sense to make occasional inferences about the world.

---

[20] See my "Neuroethology and the Philosophy of Cognitive Science," *Philosophy of Science* (Proceedings), LXVII (2000): S404-17.

[21] See my "Fixing Content and Function in Neurobiological Systems: The Neuroethology of Electroreception."

At this point, I can finally address an account that has likely bothered some readers by its absence in this discussion: ecological psychology. In *The Senses Considered as Perceptual Systems*, J. J. Gibson[22] offers a characterization of the senses at odds with many other accounts in psychology. As Heil points out (*op. cit.*, pp. 10-11), Gibson does not directly address Aristotle's and the star-nosed mole problems, but such an account seems implicit in his work (and I shall rely on Heil's explication here). According to Heil, "Gibson's fundamental notion is that perceiving is the picking up of information about the world made available to the perceiver by various sorts of physical stimulation" (*op. cit.*, p. 10). The notion of "information" is critical to the Gibsonian account, and it is at the crux of why I part company with them. Gibsonians hold that it makes sense to attribute a sensory modality to any organism that can act on structured stimuli of a particular physical type *regardless of how that information is obtained by the organism*. Take the case of a blind person equipped with a video camera and mechanism that converts the visual image into an isomorphic pattern of vibrating pins placed against the skin, as in the 1970s experiments with "tactile visual substitution systems" (TVSS) made famous by the work of P. Bach-y-Rita.[23] On the ecological psychology account, therefore, it makes sense to say that TVSS-equipped persons can see (albeit poorly)—that they have a visual modality.

Given what I just said about the distinction between detection and reception, however, it should be clear why I disagree here. A TVSS-equipped, but otherwise blind, individual is capable of visual detection, not visual reception. Such persons should no more be said to have a visual modality than all of us should be said to have an electrical modality just because we can detect electricity with our tongues. It is true that they are getting visual information about the world, but they are getting that information via their tactile modality. Giving a blind person a TVSS system does not give them a modality they did not have before. Rather, it allows them to make better use of the modalities they already have. To the extent that ecological psychology fails to draw this distinction, it does not give us an adequate account of the senses.

---

[22] Boston: Houghton Mifflin, 1966; see also *The Ecological Approach to Visual Perception* (Boston: Houghton Mifflin, 1979). Noë (*op. cit.*) also defends a Gibsonian account of the senses.

[23] *Brain Mechanisms in Sensory Substitution* (New York: Academic, 1972).

The distinction at play here is that between the content of the senses and the mode of perceptual interaction between the organism and the world. Sensory modality is not simply an issue of what things in the external world can become the content of an individual's psychological states, but rather the mode by which that content comes into the organism. It may well be the case that a blind person can come to have every propositional attitude a sighted person has. But such an individual is still blind; he lacks the modality of vision. He has one less modality than typical members of his species. He has just cleverly jury rigged a sensory system dedicated to the reception of mechanical distortion (his skin) into one capable of providing him with generally reliable information about the electromagnetic spectrum. But his perceptual mode of interaction with the "visual world" is tactile.

What this example shows is that there is more that we may wish to learn about perception than simply what modalities are at play. But an understanding of modalities will help us recognize an important way in which a TVSS-equipped blind person and others differ perceptually: one (the blind person) lacks a sensory system dedicated to the perception of electromagnetic stimuli which another person (a sighted one) has. One lacks a modality that the other one has, just as electric fish possess a modality (electroreception) that humans lack, although, thanks to our ingenuity, we can sometimes obtain information about the electrical world.

IV. TWO CRITERIA TO BE REJECTED

In the preceding section, I argued for the four criteria that I believe are necessary to differentiate the senses. Other criteria have been proposed. On my account, however, they are misapplied to Aristotle's and the star-nosed mole problems. Nonetheless, they are worth considering in some detail, if only because they have a great deal of intuitive plausibility.

In perhaps the twentieth century's most cited discussion of Aristotle's problem, Grice (*op. cit.*) discusses a set of four criteria for distinguishing the senses: proper objects, sensation, neurobiology, and physics. The bulk of Grice's paper, however, is *not* spent arguing for the adequacy of these four criteria, but rather discussing the relationship between the first two:

(5) *The proper objects of sensation*: the special features detectable by the operation of the senses; that is, by "the differing features that we become aware of by means of [the senses]" (*op. cit.*, p. 250). For example, through vision, we become aware of colors; through audition, degrees of loudness.

(6) *Sensation*: "It might be suggested that two senses, for example seeing and smelling, are to be distinguished by *the special introspectible character* of the experiences of seeing and smelling; that is, disregarding the differences between the characteristics we learn about by sight and smell, we are entitled to say that seeing is itself different in character from smelling" (*op. cit.*, p. 250).

Grice's first criterion (our (5)) is perhaps the most venerable approach to the issues at hand. This is the account to be found in Aristotle's *De Anima*, Book ii.[24] He points out that there are qualities, such as color, that are perceived only by a single sense. These are the "proper objects" of these senses, "the things to which the very being of each sense is naturally related" (*ibid.*, p. 56). Aristotle goes on to contrast the proper objects with the "common sensibles"—number, shape, magnitude, motion, rest, and so on—which are qualities that are sensible by multiple modalities. Therefore, on this account, one first identifies the formal objects that are proper to only one sense, and based on that list we can derive a classification of the senses related to those objects.

In a sense, Aristotle's criterion can be seen as a primitive version of the physics criterion (1) in that a proper object account draws a strong connection between the categories of physics and those of our phenomenology. The categories of Aristotelian physics are the categories of common-sense physics. Of course, it is always open to us to reinterpret Aristotle's account in light of contemporary physics and argue, for instance, that "wavelength of electromagnetic energy" and not "color" is the *true* proper object of vision. But such a reinterpretation represents such a significant change to the common-sense nature of Aristotle's original proposal that it seems appropriate to identify it as a *different* proposal.[25]

---

[24] According to Richard Sorabji's reading of Aristotle, which I follow closely in this discussion—"Aristotle on Demarcating the Five Senses," *Philosophical Review*, LXXX (1971): 55-79.

Grice's[26] second criterion (our (6)) proposes to individuate modalities by reference to the "*special introspectible character*," or what might be more commonly referred to as the distinct *sensations* or *qualia* associated with each given sensory modality. Since each sense has its own unique experiential quality, we can use these qualities to differentiate the senses. The second criterion would have us catalogue the different experiences we have, sorting them in terms of similarity and difference, and end up with several sets of related experiences that are the different modalities: the visual experiences, the olfactory experiences, and so on.

While these two criteria seem independent of one another, one of the central goals of Grice's article is to argue that they are not: any attempt to make suggestion (5) work leads to difficulties which seem soluble only if we bring in suggestion (6); yet suggestion (6) in turn involves difficulties which seem soluble only by adopting suggestion (5) (*op. cit.*, p. 259). I do not wish to regurgitate Grice's clear, and to my mind correct, arguments for the interdependence of these criteria. But to motivate my critique of the sensation criterion below, let me present half of it—the dependence of proper objects on sensations.

---

[25] At this point, I should mention two relatively recent approaches that are suggested improvements on Aristotle's "proper object" account. Both are attempts to spell out an account of the senses based on the content of perception; spelled out in terms of either "key features" (Roxbee Cox, *op. cit.*) or "primary objects" of perception (David H Sanford, "The Primary Objects of Perception," *Mind*, LXXXV (1976): 189-208). In addition, Lopes (*op. cit.*) has recently endorsed a sensation-based account. He does not actively argue for it, however. Instead, he concentrates his attention on showing the inadequacies of Dretske's position (*Naturalizing the Mind* (Cambridge: MIT, 1995)) that we can make sense of the senses using an appropriate theory of content At best, Lopes seems to ignore the possibility of unconscious sensory modalities. At worst, he rules them out by stipulation: "what it is like to perceive in one sense modality is different from what it is like to perceive in others—each has a unique 'phenomenal character'—and this is a fact which any theory of perception must take account" (*op. cit.*, p. 439). My disagreement with this claim should be clear from the discussion of criterion (6) below. (Cf Dretske, "Reply to Lopes," *Philosophy and Phenomenological Research*, LX (2000): 455-59. See also Nelkin (*op cit.*), and Ross (*op. cit.*) for related content-based approaches to the senses that regrettably I do not have space here to engage.)

[26] Grice does not use the terms 'proper objects of sensation' or 'sensation' or 'qualia' in his discussion of the senses. But what he says about each position seems to line up well with the common uses of these terms, so for pedagogical reasons, and to place Grice's discussion in a broader context, I shall use these terms to identify his criteria.

K.P. - 1373

Grice begins by following the suggestion of the proper objects criterion. He identifies a series of perceptual features of the world that are independent of the agent—for example, color, pitch, and temperature. This then allows us to individuate the senses using this list, without any reference to the way such features are experienced by a perceiver. At this point, a problem arises: "According to [the proper object criterion], certain properties are listed as visual properties, certain others as tactual properties, and so forth; and to say that color is a visual property would seem to amount to no more than saying that color is a member of a group of properties the others of which are... [*sic*]. This leaves membership of the group as an apparently arbitrary matter" (*op. cit.*, p. 255). That is to say, relying on the proper objects of sense does not tell us by virtue of what these properties are the proper objects of *vision*, whereas those properties are the proper objects of *touch*. Of course, the obvious thing that classes these properties together is that we *see* the visual ones, and *tactually feel* the tactile ones. But to invoke this feature is to revert to the sensation criterion.

A second problem with the proper objects criterion is that it breaks down once you try to put it into practice. That color is a proper object of vision seems uncontroversial. But is warmth a proper object of the tactile sense? This seems correct until we realize that one can occasionally see the temperature of objects, as when the blacksmith sees that the red glowing metal bar next to the furnace is hot. The clear response to this worry is to draw a distinction between *directly* sensing temperature (as the tactile sense does) and *inferring* temperature (as we sometimes do with our visual senses). The *proper* objects of a sense are those which it directly senses. Grice demonstrates over several pages, however, that we cannot make sense of this suggested distinction without cashing out the notion of "directness" in terms of having a particular qualitative experience; for example, directly sensing warmth is to experience a sensation of warmth, something you get through the tactile sense and never through the visual sense (*op. cit.*, pp. 251-55). Once again, criterion (5) only makes sense by invoking criterion (6).

Given the above arguments, the cogency of criterion (5) rests on the foundation provided by criterion (6). How firm is this foundation? Grice argues that any account of the senses in terms of experiential character of sensation rests, in turn, on an account in terms of proper objects. I want to take a different tack. I accept that a proper object account rests on an account of sensation, but argue that we cannot use the experiential character of sensation to differentiate the senses. *Both* proper objects and sensations are nonstarters when it comes to solving Aristotle's and the star-nosed mole problems.[27]

What is wrong with using the experiential character of sensation as a criterion for individuating the senses? There are two problems with this approach. First, while potentially useful for solving Aristotle's problem, it is less clear what use it is in solving the star-nosed mole problem. Do we have to believe in the existence of electrical qualia, say, before we can sensibly talk of an electrical modality in electric fish or decide whether the star-nosed mole has an electrical modality? It seems to require that we be able to answer Thomas Nagel[28] type questions concerning what it is like to be an electric fish or star-nosed mole. This point echoes one made by Coady: "Further difficulties will arise for a Gricean view from the fact that we commonly attribute sight, touch, hearing, etc. to dumb animals and here we not only make no use of [the sensation criterion] but there seems to be no way in which we could" (*op. cit.*, p. 111).

Of course, there is never a guarantee that philosophical analysis will make science easy (or even possible), and we might just have to accept that science cannot answer the star-nosed mole problem until it has overcome the worries raised by Nagel. I think, however, that there are other reasons for rejecting criterion (6) which render Nagel mute on the issue at hand. Grice's proposal runs into problems even with humans, because there is reason to believe that there are legitimate sensory modalities that *lack a special introspectible character altogether.*

Consider the case of the vomeronasal system. Admittedly, there is still controversy as to whether humans possess this modality, but over the past decade evidence in its favor has begun to mount. Furthermore, if we in fact possess this system, two things about it are striking: first, it plays a significant role in human behavior; and, second, we experience no sensations associated with this modality—there is no "special introspectible character" here, hence no basis to individuate this modality from any other. It would appear to be a modality without sensory experiences.

In almost all vertebrates investigated, airborne chemicals are detected by multiple anatomical systems. One is the well-known system involving the olfactory epithelium within the nasal cavity containing chemosensory cells that project to the olfactory bulb. In humans, this is the system responsible for smell and taste experiences. In most animals, this system is primarily responsible for the detection and evaluation of food.

---

[27] In addition to Grice, Mark Leon—"Characterising the Senses," *Mind and Language*, III (1988): 243-70—also argues for a sensation-based approach to differentiating the senses. But he is mainly concerned to counter the criticisms of Grice and to distinguish his account from belief-based approaches. He does not address the issues I raise here. Lopes (*op. cit.*) also endorses criterion (6), but see footnote 25 above.

[28] "What Is It Like to Be a Bat?" *The Philosophical Review*, LXXXIII (1974): 435-50.

There is a second system, however, that is primarily social in function. The vomeronasal organ is located in a pair of pits on either side of the nasal septum. The vomeronasal system is primarily responsible for detecting pheromones, which in turn have been shown to play a central role in reproductive behavior.[29] For example, animals with lesioned vomeronasal organs typically exhibit greatly reduced sexual behavior. Similarly, artificially stimulating this organ and the nuclei to which it projects generally produces sexual behaviors in the animal so manipulated, even in the absence of appropriate conspecifics.

Do humans have a vomeronasal system? The textbook answer has traditionally been that while this system is present in human fetuses, it disappears during normal development.[30] In recent years, however, this received wisdom has been called into question.[31] In terms of anatomy—contrary to the textbooks—vomeronasal pits are present in most adult humans.[32] Furthermore, the pits are innervated by sensory neurons.[33]

Behaviorally, there is a growing list of findings in humans that closely resemble behaviors carried out by the vomeronasal system in other species. This, combined with the fact that these behaviors seem to have no conscious correlates, suggests the presence of a nonconscious modality in humans. First, it has been reported that individuals can detect the gender of another based on smelling the breath alone. Some women are reportedly able to identify the gender of a breather with an accuracy of ninety-five percent![34] Second, clinicians have observed that damage to the nerves in the nasal region is often, but not always, associated with a loss of interest in sex. (Because medical students have typically been taught that humans do not have a vomeronasal organ, however, nobody has pursued a study of naturally occurring lesions to the vomeronasal versus olfactory epithelial systems in humans. Therefore, to my knowledge, no attempt has been made to try to tease apart the functions of these two systems, as has been done in nonhuman animals.)

---

[29] M. Halpern, "The Organization and Function of the Vomeronasal System," *Annual Review of Neuroscience*, X (1987): 325-62.

[30] Elizabeth C. Crosby and Tryphena Humphrey, "Studies of the Vertebrate Telencephalon," *Journal of Comparative Neurology*, LXXI (1939): 121-213, especially p 121.

[31] Robert Taylor, "Brave New Nose: Sniffing Out Human Sexual Chemistry," *Journal of NIH Research*, VI (1994). 47-51.

[32] A. Johnson, R. Josephson, and M. Hawke, "Clinical and Histological Evidence for the Presence of the Vomeronasal (Jacobson's) Organ in Adult Humans," *Journal of Otolaryngology*, XIV (1985): 71-79.

[33] S. Takami, M. L. Getchell, Y. Chen, L. Monti-Bloch, D. L. Berliner, L. J. Stensaas, and T. V. Getchell, "Vomeronasal Epithelial Cells of the Adult Human Express Neuron-specific Molecules," *NeuroReport*, IV (1993): 375-78.

[34] Rivlin and Gravelle, p. 154.

While the science is admittedly controversial here, the possibility of a human vomeronasal system stands as a potentially interesting case of a modality without a special introspectible experiential character. Women who can guess the gender of breath do not report that they experience "male" versus "female" qualia associated with the breaths. Indeed, subjects are generally surprised to be informed that they are so good at distinguishing the smells. Gender detection via the putative vomeronasal sense seems akin to an olfactory version of blindsight.[35] And if it is a modality that lacks qualia, then criterion (6) cannot even begin to distinguish this modality from others.

Many will no doubt be surprised by the above arguments that the differentiation of the senses is independent of any appeal to the experiences associated with the senses. A commentator[36] on an early version of this article put it this way: "Much of the bad press over qualia is well-deserved; but if there is one place experiential qualities have a safe home, I would've thought it would be with the sense modalities." In response, I would say that experiential qualities do have a safe home in the modalities. Much of what we experience, we experience through the senses. My argument here is *not* that there is no such thing as sensory experience, but rather that we should not use those experiences to differentiate the senses. Experience is often associated with the senses, but its nature does not define the difference between the individual senses.[37]

---

[35] L. Weiskrantz, *Blindsight: A Case Study and Implications* (New York: Oxford, 1986).

[36] Tom Polger made this comment at the 1999 *Society for Philosophy and Psychology* meeting in Palo Alto, California.

[37] For reasons of space, one issue I have not discussed here is the assumption that the senses are "fairly discrete systems," as Leon puts it (*op. cit.*, p. 245, footnote 1). It is common to assume that the different senses are significantly separate and independent of one another. Indeed, this assumption motivates the questions discussed here. To my knowledge, this ubiquitous assumption has received even less attention in the philosophical literature than the star-nosed mole and Aristotle problems; a curious observation given the existence of empirical phenomena that raise questions about it. I am thinking here of synesthesia: Richard E. Cytowic, *Synesthesia: A Union of the Senses* (New York: Springer, 1989), *The Man Who Tasted*

V. IMPLICATIONS: WHAT SENSES DO HUMANS AND OTHERS HAVE?

The proof, they say, is in the pudding, and I conclude by showing how my account deals with some actual cases by spelling out its philosophical implications. First, let us consider Aristotle's problem. On my account, vision, hearing, touch, smell, and taste come out as different senses. Each involves sense organs dedicated to the detection of a different class of physical stimulation. If future science pans out as I have described it above, humans should be said to have a vomeronasal modality as well. We also have a vestibular sense (based in the semicircular canals in our heads) as well as a proprioceptive system (based in the stretch receptors in our skeletal muscles).

At the same time, on my account, humans should not be said to have an electric sense because we have as yet discovered no organ that is dedicated to the processing of electrical information in our environment. As noted above, we can access electrical information through our tongues, but only by electrically stimulating sensory systems that are normally responsive to other physical properties—physical properties that are part of the normal human sensory environment. On the other hand, electric fish *should* be said to have an electrical sensory modality because these organisms have organs that have evolved specifically to process biologically meaningful electrical stimuli in their environments.

What about the star-nosed mole? On my account here, the nose of a star-nosed mole would seem to be best thought of as mediating a tactile sense. Although it is true that sensory cells in the nose of this animal can be stimulated by the presentation of electricity, as with the case of the human tongue, I realize that this fact does not settle the issue. All sensory cells, when blasted with enough electricity, can be made to respond. The question is whether there is any reason to believe that the amount of electricity required to stimulate the nose of a star-nosed mole is within the range that would make it likely that their noses had evolved this capacity. Is the detection of electricity a dedicated function of star-nosed mole noses? To date, such evidence

---

Shapes: A Bizarre Medical Mystery Offers Revolutionary Insights into Emotions, Reasoning and Consciousness (New York: Putnam, 1998), and the McGurk effect—H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, CCLXIV (1976): 747-48. In the neurological-condition synesthesia, the proper objects of one sense result in experiences in more than one modality—for example, "colored-hearing." In the McGurk effect, normal humans will literally hear the same auditory speech stimulus differently, depending on what they see (lip movements of one sort versus another). It is not clear what effect such interaction between senses has on the story I am telling here. Unfortunately, to explore the issue fully would require another article.

has not been forthcoming. On the other hand, the work of Catania (*op. cit.*) indicates that these same sensory cells are responsive to tactile stimulation and, furthermore, that the range of mechanical stimulation required falls within an ecologically plausible range. With such a nose, a star-nosed mole is able to make all sorts of useful sensory discriminations of the texture, motion, and shape of objects that it places its nose upon. What is more, Catania has done careful comparative studies indicating that the sensory end organs on the nose of the star-nosed mole are likely modified versions of the tactile-sensing end organs found in related species of moles, which in turn are modified versions of the basic tactile-sensing end organs found in most mammals, including humans. All of these discoveries point to the conclusion that the nose of a star-nosed mole is properly thought of as a tactile-sensor.

The sort of story told about the star-nosed mole can be used as a template in the cases of all putative sensory modalities in nonhuman animals. If, for example, one wishes to argue for the presence of a magnetic sense in migratory birds or an electrical sense in the platypus, then the same set of evidence needs to be collected. One must characterize the target of the sense in physical terms (What range of magnetic stimulation? Exactly what electrical properties?). One must demonstrate via the organisms' behavior that the organisms in question can make use of the alleged sense. One must find a sense organ that can transduce this information from the environment to the CNS of the organism. Finally, one must demonstrate that this organ has the evolutionary or developmental function to carry out such sensory transductions. Only if you do all four of these things can you properly talk of the animal as having the sense in question.

Those are the scientific implications of my account. What of the philosophical ones? If the account that I have presented here is plausible, then it represents a strong, naturalized alternative to the more common-sense approaches to the issues typically favored by philosophers. For the purposes of the perceptual sciences, at least, distinguishing the senses from one another is not a matter of such folk-scientific entities as the proper objects of sensation or some specific qualitative feel of conscious perceptual experience. Strictly speaking, this is not to say that qualia do not exist, but rather that they do not have a role to play in this particular scientific story, however useful they may be to our folk understanding of ourselves. Defenders of qualia need to look elsewhere for scientific legitimacy.

If you are unconvinced with the positive story I have told here, I hope to have at least persuaded you that there are interesting philosophical and scientific questions yet to be answered concerning the

THE MAXIMS PROBLEM*

As ethical formalists see it, human agency is free and principled.[1] They therefore think that the conduct and character of free agents can always be evaluated by their principles: either by formal features of these principles, like consistency or universality, or by the coherence of principles with the form or structure of free agency in general. Yet someone might object to a formalist appraisal of his conduct, claiming that he rarely acts on principle.[2] He might even explain that he avoids principled conduct because he finds that commitments to general principles can constrain his freedom to respond to the particulars of practical experience. But formalists would easily dismiss this line of objection, explaining that, as they see it, acting on principle is a *logical* requirement on free agency.

When formalists claim that human agency is free and principled, they do not mean that we are free to act on principle if we choose. They mean that whenever we act by our own free choice, we act on a principle—whether we like it or not, even whether we recognize it or not. For they see a *necessary* or *conceptual* connection between free choices and the principles that Kantian formalists call "maxims." So they think it impossible to act freely without a maxim, or to choose to go unprincipled.[3] By insisting on this logical link between free choices and principles, therefore, formalists can guarantee that all freely chosen actions will be subject to moral evaluation as they conceive it. This is the primary, theoretical benefit they derive from linking freedom and principled agency conceptually. But the full cost of this conceptual connection seems underestimated. This, in general terms, is the *maxims problem*.

---

* I wish to thank Jane Kneller for comments on an earlier draft of this paper, and James Maffie for valuable criticism.

[1] Formalist approaches to ethics include all those basing moral evaluation of actions on formal features of agents' personal or subjective principles of action. While Immanuel Kant and most of his followers are leading formalists, the "universal prescriptivism" of R. M. Hare should qualify as a non-Kantian formalist approach to ethics—*The Language of Morals* (New York: Oxford, 1964). Also included among recent formalists are Thomas Nagel and Alan Gewirth.

[2] Cf. Gerald Dworkin, "Unprincipled Ethics," *Midwest Studies in Philosophy*, xx (1996): 224-39; and Joseph Margolis, *Life without Principles* (Cambridge: Blackwell, 1996), pp. 146-47, 150.

[3] Hare's variant on the Kantian view of agency is that, although acting without principles is possible, it is a fact that no one acts without a principle (*op. cit.*, pp. 56-59)

0022-362X/02/9901/29–44

To begin exploring the difficulties with the formalists' view of maxim principles in free agency, consider two assumptions popular among those recently developing the Kantian tradition. The first is that '*Choosing is adopting a maxim*'.[4] This assumption elegantly preserves the conceptual connection formalists require between free choices and principles. In addition to guaranteeing that there will be a maxim principle corresponding to every freely chosen action, it also helps explain the sense in which maxims are held to be "subjective" principles: principles internalized by, or belonging to, the agents who freely act on them.[5] The second assumption is that '*Maxims are prescriptions*'.[6] Many formalists think of maxim principles as statements of policy, and suppose that to act on one's adopted maxim is simply to do what it says to do, or to obey its command.

But if we now join these two assumptions, the result is a two-step model of agency that works as follows: whenever we choose, we at least implicitly adopt a maxim principle prescribing how to act; then we subsequently act on our adopted principle by doing what it tells us to do. Yet the difficulty arises at the beginning of the second step. For it is a fact about human agency that we typically cannot do what we are told to do, not even what we have commanded ourselves to do, without choosing to do it: our first response to any prescription is always to choose whether to obey it or not. So once we begin by adopting a principle telling us what to do, we must next choose whether to do it. And either way we choose, we shall once again adopt a maxim; for choosing is adopting a maxim. Then, after adopting a maxim for the second time, we shall be required again to respond to a principle telling us what to do; for maxim principles are prescriptions. It should therefore be clear that, if choosing really is adopting a maxim, and if maxim principles really are prescriptions, then human "agency" is an endless cycle of choice and prescription which can never terminate in action.

---

[4] Henry Allison implies that Kant and Kantians assume choosing is adopting a maxim in an "act of incorporation" when he explains that "we think of reason as determining the will by legislating to it, that is, by providing the laws or principles...which govern, without causally necessitating, its acts of incorporation or, more simply, its choices"—*Kant's Theory of Freedom* (New York: Cambridge, 1991), p. 51. But others seem not to agree completely that a choice is a maxim adoption. Barbara Herman may imply here that some choices to act are maxim adoptions, while others are not: "deciding to act in a certain way *can itself be* subscribing to a principle of action"—*Morality as Rationality* (New York: Garland, 1990), p. 40. Cf. also Hare, p. 59.

[5] The subjectivity of maxim principles is important because particular actions can usually be covered by any of several general principles And as most formalists see it, agents should be evaluated only by the specific principles which actually guide their conduct, or which they acturally choose, subscribe to, or somehow make their own. Maxim subjectivity seems also to be part of the best solution to the "problem of relevant act descriptions" which has occupied the attention of Kant's recent interpreters. See Onora Nell, *Acting on Principle* (New York: Columbia, 1975), pp. 12-42, 137-43; and Herman, *The Practice of Moral Judgment* (Cambridge. Harvard, 1993), pp. 143ff.

[6] Kant unfortunately made it difficult to tell whether he thought maxims prescribe action. Allison implies that they do in describing a maxim as "a rule or principle of action, which dictates that we ought to pursue the satisfaction of [an] inclination or desire..." (*op. cit.*, p. 40, cf. also pp. 89-90). The assumption that maxims are prescriptions seems also to be at the heart of the theory of agency grounding Hare's universal prescriptivism. Christine Korsgaard implies that maxims are prescriptions as well with her unique characterization of a (good) maxim as an *intrinsically normative entity*—*Sources of Normativity* (New York: Cambridge, 1996), p 112.

Formalists need to maintain a conceptual connection between free choice and maxim principles; otherwise, agents might freely opt out of acting on principle. Most therefore maintain this conceptual connection by assuming that choosing is an implicit operation on a maxim principle: something like adopting or subscribing to it. But what few have noticed is that connecting choices and principles in this way places restrictions on the form and function of maxim principles in human agency. For if choosing is adopting a maxim, then maxims cannot take the form of prescriptive rules or policies prompting us to choose. Of course, there are probably other, more suitable forms that maxim principles can take. Formalists need not assume that the subjective principles of human agency are prescriptive, and some do not. There may also be other ways formalists can maintain the conceptual link between choices and principles which do not require the assumption that choosing is adopting a maxim. In what follows, I explore what seem to be the most promising alternatives.

Just below, in section I, I consider alternatives to the second of the two assumptions above, that maxim principles are prescriptions. This involves reconceiving maxims to function in agency more like agents' formed intentions than internalized policies or imperatives. But a subsequent problem arises in section II, centering on the generality of nonprescriptive principles of agency. Section III then focuses on possible responses to this problem which involve revising the first assumption that 'Choosing is adopting a maxim'. More expansive conceptions of choosing are considered which identify choosing with not just the one operation of adopting a maxim, but with a plurality of operations on maxim principles. Although this strategy seems helpful, it will unfortunately always sever the conceptual connection formalists require between free agency and principles. In section IV, I critique a recent model of practical rationality likely to seem exempt from the maxims problem. After finding otherwise, I conclude by suggesting a couple of ways formalists can maintain that human agency is free and principled without modeling it as an endless cycle of choice and prescription. As a class, maxim principles can be suitably understood either as expressions of specific intentions in acting, or as representations of fixed dispositions of conduct. Each suggestion can adequately sustain the conceptual link formalists require between actions or choices and principles. But each also significantly constricts either the scope of free agency's principles, or the autonomy of principled agency. For these reasons, it is unlikely that either of these concluding recommendations will be happily adopted as a solution to the maxims problem.

I. MAXIM MODALITY

The assumption that 'Choosing is adopting a maxim' enjoys the virtue of simplicity, though it may be only simplistic. It is too early at this point to be certain. But as long as we cling to that assumption, we must avoid thinking that agents' maxim principles function to guide or direct their choices. And that should lead us to agree that maxim statements are poorly expressed as prescriptions. On the other hand, formulating maxims in the imperative mood might be acceptable so long as they are not supposed to function in agency as ordinary, choice-prompting commands. Perhaps some formalists favor the imperative mood for expressing maxim statements because they think of maxims not as self-commands to action, but as something like "internal directives" of agency.[7] An agent having adopted or accepted a maxim principle is usually thought to have internalized a reason for action. So stating maxims in the imperative mood may be nothing more than a convention for distinguishing practical reasons from theoretical reasons, or reasons for acting from reasons for believing.[8]

In that case, however, maxim principles seem better expressed in the infinitive mood, if only because this avoids confusing them with choice-prompting commands. For adopting an internal directive as a reason for action seems comparable to forming an intention. If I adopt the imperative formula 'Borrow money you know you cannot repay!' as a reason for acting, then I seem simply to intend 'to borrow money I know I cannot repay'. That prescription seems to be my principle, or I seem to have subscribed to it, just in case I have formed the intention expressed by the infinitive. And this means that I need not subsequently choose to act on my adopted principle, for I have already chosen to act on it in forming my intention. Thus, if choosing is adopting a maxim, and if adopted maxims are properly expressed by infinitives, then once a maxim is adopted it should not be necessary to choose again in order to act on it.[9] Alternatively, maxims might be properly expressed in the indicative mood. The content of someone's lying-promise maxim might be stated simply as, 'Borrowing money I know I cannot repay'; or even with a declarative statement like, 'I borrow money I know I cannot repay'.[10] And if maxim principles are indicatives, then choosing or adopting a maxim seems to entail making it a true description of oneself. In that case, therefore, our maxim principles would always express just how we act, or how we have determined ourselves to act. But they would never express how we ought to act, nor command how we should choose.

---

[7] There is ample reason to think of the prescriptive principles driving Hare's theory of agency in this way; for he maintains that subsequent to sincerely "accepting" or "assenting to" such a prescription there is no question of the agent's then choosing whether to act on it or not (*op. cit.*, p. 20).

[8] Cf. Nagel, *The Possibility of Altruism* (Princeton: University Press, 1970), p. 65.

II. MAXIM GENERALITY

A more satisfactory model of formalist agency seems to result when maxim principles are expressed in infinitive or indicative moods: a model that does not always generate an endless cycle of choice and prescription. But if so, then the action scope of maxim principles should always be singular.[11] That is, the numerical relation between agents' maxims and their actions should always be one-one, so that every act token has its own maxim, adopted just as the agent chooses or forms her intention to act. For if the action scope of infinitive or indicative maxim principles is ever general, so that a single maxim may cover many actions, then principled agency seems considerably less free or autonomous than formalists usually suppose.[12]

---

[9] An intriguing suggestion of Herman's seemingly along these lines is that "the maxim [represents] the way an agent wills—how, to put it somewhat dramatically, she sets herself to change the world for what she takes to be good reasons"—*The Practice of Moral Judgment*, p. 221.

[10] Cf. Gewirth, *Reason and Morality* (Chicago University Press, 1978), pp. 42-48, and *Self-Fulfillment* (Princeton. University Press, 1998), p. 81.

[11] Jan Narveson clarifies some of the multiple generality of maxim principles by distinguishing between their *agent scope* and their *patient scope*, in "The How and Why of Universalizability," in Nelson T. Potter and Mark Timmons, eds., *Morality and Universality* (Boston: Reidel, 1985), pp. 3-44, especially pp. 16-17. In a similar vein, we can usefully identify the *action scope* of a maxim principle as the range of actions it covers.

[12] A recent controversy among Kantian theorists is over maxims' *action scope*. O'Neill, Rüdiger Bittner, and Otfried Höffe are among those who maintain that all maxims are principles expressing general rules or policies; cf. O'Neill, *Constructions of Reason* (New York: Cambridge, 1989), p. 129. On the other hand, Herman, Allison, and Robert Louden have countered with arguments that at least some maxims must be singular in scope. All seem nevertheless to agree that maxims can be general, and that many theoretically important maxims must be; see footnote 14, below

Imagine an agent having made it her general principle, 'Always to borrow money I cannot repay'; or as a conditional, 'To borrow money I cannot repay whenever I am in financial trouble'. In freely adopting an infinitive principle like this, it seems she would form the intention to act that way, always. Or, supposing instead that maxims take the indicative mood, it seems by adopting that general principle she would warrant a description of herself like, 'I always borrow money I cannot repay when I am in financial trouble'. Either way, it is important to notice that after adopting general principles like these, an agent should be incapable of choosing whether or not to act on them. This is for the same reason she cannot choose whether to act on her infinitive or descriptive maxims of singular scope. She should be capable of freely adopting or subscribing to any general maxim in order to make it her principle, because that is just what we are assuming freely choosing is: adopting maxims. But once she adopts any general principle, she should have no choice but to act as it indicates, again and again in relevant circumstances. Indeed, should our agent succeed in making it her principle or intention always to act a certain way, or should she actually make it a true description of herself that she acts that way whenever the specified circumstances obtain, it is hard to see how she could ever do otherwise than act on her adopted maxim in the relevant circumstances. And in that case, her life's conduct should be largely determined by the free choices of her youth.[15]

A suggestion for avoiding this problem is simply to bar maxim principles of general scope from the theory of human agency, in order to enhance agent autonomy. But the suggestion is problematic, as it raises a troubling theoretical question: Why insist on principled agency at all? The function of principles in theoretical reasoning is to justify many particular judgments. So their function in practical reasoning seems likewise to justify many particular actions. Thus, if our adopted maxims can stand only in strictly one-one relations to actions, they seem not to function in agency as principles. Saying that every particular action must have its unique principle amounts to little more than saying that we always act on our specific intentions. It justifies thinking of free agency as intentional, but it seems not to support any familiar or ordinary conception of principled agency.[14] If maxim principles must be singular in action scope, in other words, then the formalists' conception of rational agency stands in need of revision; but more of this later.

---

[15] The chief attraction of prescriptive or choice-prompting maxims is undoubtedly their immunity to such problems related to generality of action scope. For a maxim that merely *prescribes* acting a certain way leaves the agent free to choose whether to follow it in particular circumstances.

III. MAXIMS AND CHOICE

An uneasy tension seems therefore to strain the formalists' conception of principled free agency. If the subjective principles of human agency are general, then freedom or autonomy appears increasingly constricted with the adoption or acceptance of each new principle. So if our agency is to remain fully or perpetually autonomous, its so-called "principles" are not what we have thought they were; for they can evidently be no more extensive in scope than particular intentions for specific actions. Yet it may now be objected that this apparent tension between freedom and the subjective principles of agency arises only because of an unnecessarily restrictive assumption about choosing. In order to preserve the crucial, conceptual connection between free choice and principles required in formalist models of agency, it seemed convenient at first to identify choosing with a unique operation on a principle, namely, adopting a maxim. Linking choices and principles this way precludes the possibility of anyone's choosing to act without a principle, and thus guarantees every freely chosen action a corresponding principle subject to moral assessment. But there may be other, less restrictive conceptions of choosing that are suitable for achieving the same end.

A nearly obvious alternative would be that 'Choosing is either adopting or discarding a maxim'.[15] If we may choose by either adopting or discarding our maxims, then we seem to enjoy complete autonomy in managing our principles. Subsequent to adopting any general maxim, we may be determined always to act on it, but only so long as it remains our maxim. For we may also enjoy the freedom to choose by discarding any of our adopted maxims, thereby avoiding having to act on them in the future. We might also be able to re-adopt any previously discarded maxim, if convenient. Although we may not technically be capable of choosing whether to act on maxims we have adopted, we should still be free to choose to reject determining principles of our past choices. This way someone's once adopting a lying-promise maxim need not lock him into a life-career of bad-faith borrowing.

---

[14] Aside from this problem, there is the weighty concern that general maxim principles are thought to characterize dispositional features of human conduct. Kantians typically ground moral duties of virtue, or so-called "imperfect duties," on maxims covering general patterns or modes of conduct. See Herman, *The Practice of Moral Judgment*, pp. 39ff., and Louden, "Kant's Virtue Ethics," *Philosophy*, LXI (1986): 479-82. Hare's formalist ethics likewise emphasizes the generality of principles, for both "psychological and practical reasons"—*The Language of Morals*, pp. 60-61; and *Moral Thinking* (New York. Oxford, 1981), pp. 35-39 For Nagel, "All reasons are in some respect general.... If there is a reason to do something on a particular occasion, it must be specifiable in general terms which allow that same reason to be present on different occasions, perhaps as a reason for doing other things" (*op. cit.*, pp. 34-35).

[15] Cf. O'Neill: "It is a corollary of Kant's conception of human freedom that we can adopt or discard maxims..."—*Constructions of Reason*, p. 106.

Another attractive conception of free choice could be that 'Choosing is either adopting or acting on a maxim'. Contemporary Kantian moralists have often compared maxims to personal policies, to be freely adopted, as well as freely acted on in appropriate circumstances. Policy statements can be expressed as general dictates or prescriptions; and maxims might be as well, provided our conception of choosing is liberally expanded beyond the simple operation of maxim adoption.[16] For if we allow that one way of choosing in free agency is adopting a policy, while another way of choosing is acting on an adopted policy, then there is no reason why maxims cannot be aptly expressed as prescriptions. The problem with thinking of maxims as prescriptions at first was that since we then supposed adopting a maxim was the only way to choose, choosing in response to adopted prescriptions required endlessly adopting further prescriptions. But if we are also capable of choosing by acting on our adopted maxims, then we have another way of responding freely to the prescriptive personal policies we adopt.

Either of these expansive, more liberal conceptions of choosing may seem to accommodate truly free and truly principled agency: autonomy is enhanced, while maxim principles can be general. Yet it remains important to guard against overexpanding our options in choosing, to include operations that are not adopting, discarding, acting on (or whatever) a maxim principle. For formalists agree that acting on principle must remain nonoptional; and they guarantee this by insisting on a necessary or conceptual connection between agents' choices and their subjective principles of action. It has to be the case that nothing can count as a free choice if it is not an operation on a maxim principle, or if it is not, as we are now considering it, some member of a disjunctively ordered set of such operations on principles.

---

[16] O'Neill once clarified her conception of choosing by distinguishing between *adopting a policy* and *making a decision*—see *Acting on Principle*, p. 40. Gewirth's view appears similar, in that he thinks some of our choices are choices of principles or reasons, while others are choices of actions based on those reasons: "it is true of all [persons addressed by various moral and other practical precepts] that they are assumed to be able to control their relevant behaviors by their unforced choice for reasons and purposes they can [choose to] make their own"—*Reason and Morality*, p. 30.

But to avoid confusion, we must clarify the difference between a disjunctively ordered set of operations on principles which are *identical with* choosing, and a disjunctively ordered set of operations on principles which are *options available for* choosing. In the first case, if I adopt a maxim then my choice is free because adopting a maxim is a member of a limited set of operations that are freely choosing. In the second case, if I adopt a maxim then my choice is free because I freely choose the option of adopting a maxim from among the set of options available to me. But in this case, the connection between choices and operations on principles is not conceptual: for choosing can be conceived independently of any of the operations on principles which are available for choosing; and those options can likewise be conceived independently of choosing. In other words, nothing about choosing here tells us why only operations on principles are available as options for choosing. So it should be conceivable for someone to choose to act without a maxim. And nothing about any of those operations on principles requires that they can only be performed by agents who freely choose them. So it should also be conceivable to make someone adopt, discard, or act on a maxim through hypnosis or brain tampering.

Because of these problems with the second case, it seems formalists mindful of the required conceptual connection between choices and principles will prefer the first case, where choosing is understood as somehow identical with a set of operations on maxim principles. Yet even if a singular mental operation like choosing can in some way be identical with a plurality of operations on a principle, it seems free choosers must then always find themselves presented with a kind of *menu of choices* whenever they attempt to choose. The set of items on choosers' choice menus should be all of the ways of choosing available to them as agents acting on principle: choosing by adopting a maxim, or choosing by discarding a maxim, or choosing by acting on a maxim, and so on.[17] But then it should not be possible for choosers to choose in any of their various ways of choosing without first selecting it from among its alternatives. And since the selection of any way of choosing from the choice menu will itself always be a choice, it will have to be a special way of choosing that is not on the choice menu. It follows that choosing cannot be identified with a disjunctively ordered set of operations on principles without also expanding the concept even further to encompass a way of selecting or deciding that is not an operation on a principle.

---

[17] Cf. T. M. Scanlon: "our practical thinking takes place within a framework of maxims and is concerned with adopting, interpreting, and modifying these principles as well as with deciding, within the framework they provide, whether we have sufficient reason for acting in particular ways"—*What We Owe to Each Other* (Cambridge: Harvard, 1998), p. 53. The rational agent Korsgaard describes seems furnished with a relatively opulent choice menu, allowing additionally for the selection of multiple, and evidently prescriptive "practical identities"; see footnote 22, below.

To illustrate, imagine I believe that by my past adoption of a nonprescriptive, general maxim principle, I have determined myself always to escape financial trouble by borrowing money I cannot repay. But I am lately burdened with regret over having made this dishonest principle my policy, and I wish that I could discard it; or if the maxim must remain my policy, then I wish at least that I might be free to choose not to act on it. And now suppose I learn that all along I have been free to choose not only which principles to adopt, but also whether to act on my adopted principles. I become aware that I have what Alan Gewirth recently explained I must have in order to be free: both "desire-autonomy" and "behavior-autonomy."[18] But surely this expanded set of ways I can choose must also come with yet a third way of choosing between them. For sometimes I shall need to decide whether I shall exercise my behavior autonomy, and choose whether to act on one of my adopted principles, or whether I shall instead exercise my desire autonomy, and choose whether to revise my set of principles. How, indeed, could I enjoy both of these distinct autonomies unless I have a way of freely choosing between them? But if the two ways of choosing available to me are each operations on principles, then my third way of choosing between them is not: to choose in the third way is neither to adopt or discard, nor to act on or refrain from acting on, a principle.

This shows why choosing cannot be identical with any disjunctively ordered set of operations on principles. So unless formalists assume that choosing is nothing other than adopting a maxim, they have no basis for insisting that all my freely chosen actions are principled, whether I recognize it or not.

---

[18] "Desire-autonomy bears on the process whereby the rules are formed, it involves that...one has chosen the rules in a self-critical way...." Behavior-autonomy, "goes beyond desire-autonomy in that it bears upon one's action following one's desires or choices [of rules, aspirations]. It is not enough for the self to control its desires [rules, aspirations]...the self must also be able to control its actions in accord with those desires"—*Self-Fulfillment*, pp. 34, 36.

IV FREEDOM AND REASONS

In ancient times, Aristotle modeled human agency with a practical syllogism. Action conclusions were presented as following from two kinds of premises: major premises containing principles, and minor premises representing perceptions or awareness of circumstances. Much the same model of rational agency is endorsed by formalists of all types.[19] The syllogistic model of practical thinking explains human behavior only because of the necessity linking premises and conclusions. So a logical gap inserted between practical premises and action, to be filled by chance or choice, would result in agency that is no longer rational because the premises fail to justify or explain the conclusion. Formalist practical reasoning accordingly provides a poor model for human agency unless principles, together with circumstantial perceptions or judgments, logically entail actions, or at least intentions to act.

But if actions or intentions must always follow from practical premises by necessity, and not by free choice, then some explanation seems required in order to reconcile freedom and practical reason. For how can our actions be voluntary if they are strictly determined by principles? A long tradition of philosophers since Aristotle has distinguished between free, rational determination by principles, and unfree, causal determination by internal compulsion or external force. Only our actions following rationally from our principles are called free or voluntary, according to this tradition. And thus for these thinkers, 'Choosing is being determined to action by principles'. Once compulsion and force are excluded, it usually does not matter how the principles determining our actions are acquired or internalized, as far as freedom is concerned.[20] For theorists in the Kantian tradition, on the other hand, practical freedom or autonomy requires that we also freely adopt the principles determining our actions: 'Choosing is adopting a maxim'.

Yet general principles often conflict, as almost everyone recognizes. And when they do, rational agents must decide or choose between alternative rational determinants of their conduct. With such conflicts in mind, therefore, modern formalists have modified the traditional syllogistic model to introduce an intermediate stage of practical thinking. That is, general principles, together with informa-

---

[19] It is worth noting that just prior to Kant, Christian Wolff had used the term 'maxim' as a name for the major premise of a practical syllogism.

[20] Exceptional cases often involve fantastic and malicious principle acquisitions through neural manipulation. Examples of emotional trauma may yield more realistic cases where acquired principles lead to behavior that most will call unfree.

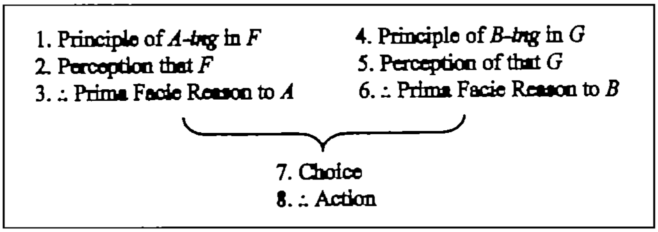

1. Principle of *A*-ing in *F*
2. Perception that *F*
3. ∴ Prima Facie Reason to *A*
4. Principle of *B*-ing in *G*
5. Perception of that *G*
6. ∴ Prima Facie Reason to *B*
7. Choice
8. ∴ Action

Figure 1

tion about relevant circumstances, are in modern times thought to yield not action conclusions, but only *prima facie reasons* for action.[21] Actions are supposed to follow ultimately upon the rational agent's freely choosing to act on one or another of her available prima facie reasons. This "extended syllogistic model" of agency constitutes a richer framework of principles, reasons, choices, and actions. For the actions it generates can be understood to follow from agents' choices of (prima facie) reasons, while those reasons are in turn thought to follow from general maxim principles. So, on this model, our internalized general principles may perpetually dictate our prima facie reasons for action as circumstances arise, while we continue to enjoy full autonomy in selecting among our available reasons for action. Hence, 'Choosing is endorsing a reason for action'.

The extended syllogistic model of agency may seem to merit consideration as a response to the maxims problem. Yet more focused attention to the model's presuppositions reveals a difficulty. As noted above, the syllogistic model explains action through the necessity linking premises and conclusions; and this feature of the model should not be thought to change with the simple addition of several more premises. But the modern and richer model interrupts the transition from practical premises and their ultimate conclusion in action, by interposing the logical "gap" of a free choice. And this must mean, despite contrary appearances, that actions presumably explained by the modern model are always underdetermined by rational principles. Action outputs of the model may always be entailed and thus explained by chosen prima facie reasons, to be sure; and all prima facie reasons available for choice may be entailed by higher principles plus perceptions or awareness of circumstances. Yet there seems to be no reason or practical premise supporting the crucial choice to act on one available reason rather than another. The model does not now provide a reason to respond to any available prima facie reasons for action at all.

---

[21] See Nagel, pp. 49-51; and Hare, *Moral Thinking*, pp. 38ff. Korsgaard treats reasons for action as if we must at least sometimes choose between them by comparing their weight, in the standard manner of those who employ prima facie reasons in explaining human conduct—see *Sources of Normativity*, p. 140.

It should suffice to say that the extended practical syllogism of modern times offers no adequate response to the maxims problem if it fails as a model of human agency on its own terms. And the model cannot bear its own burden to explain action through practical reasoning, unless we supply additional "preference principles," or higher-order maxims whose function it is to rank prima facie reasons for actions by their choice worthiness.[22] But once preference principles become internalized, as they must be in order to be effective in agency, the extended syllogistic model becomes only a slightly richer representation of its ancient ancestor.

Here, in abstraction, is the difficulty: suppose it is my adopted maxim 'Always to *A* in circumstance *F*'; while it is also my maxim 'Always to *B* in circumstance *G*'; and moreover, I perceive that I am simultaneously in circumstances *F* and *G*, but I cannot *A* and *B*. According to the extended model, my two maxim principles plus my awareness of my compound circumstances will yield both a prima facie reason to *A*, and a prima facie reason to *B*, between which I must now choose. But the problem is that none of my principles can justify or explain my choosing *A* over *B*, or *B* over *A*. Hence, if I ultimately *A*, I shall not arrive at that result by any operation of practical reasoning; and this is true despite the fact that I even had a reason to *A*, justified by one of my own maxim principles. Missing from this model of agency is a choice-justifying preference principle, or a third maxim like 'Always to *A* instead of *B* when in circumstances *F* and *G*'. But once this preference principle is adopted or internalized, each of my two earlier maxims, 'To *A* in *F*' and 'To *B* in *G*', should be inapplicable to my present circumstances.[23] So neither should be capable of justifying any prima facie reasons for action in this instance. Indeed, there should be no role at all for prima facie reasons in rational agency once conflict-resolving preference principles are admitted to the model.[24] And with the loss of alternative prima facie reasons to choose, the modern syllogistic model of practical rationality can offer no interesting response to the maxims problem. Formalists committed to this model will encounter a version of the maxims problem when prompted to explain whether choosing is adopting a maxim, or adopting a preference principle, or acting on a preference principle, or on a maxim endorsed by a preference principle.

---

[22] Nagel calls these "combinatorial principles" (*op. cit.*, p. 133). Hare locates "second-order" principles on a level of reflection called "critical moral thinking"—see *Moral Thinking*, pp. 41, 60-61. In Korsgaard's recent expansion of the Kantian model, it seems that what she calls "contingent practical identities" serve as freely adopted, second-order principles of agency, functioning to prescribe the adoption of first-order principles: maxims or prima facie reasons. In this sense, practical identities seem to be "principles of choice" that dictate whether to endorse a desire or other incentive as a reason: "Practical conceptions of our identities determine which of our impulses we will treat as reasons"—*Sources of Normativity*, p. 129. Even so, sometimes Korsgaard implies that we may subsequently choose not to act on our adopted first-order reasons, p. 104.

[23] Preference principles might be conceived to operate differently: as principles justifying choices between competing lower-order reasons for action. That is, my

IV CONCLUSION

Principled free agency, as understood by Kantians and other formalists, requires a conceptual connection between free choice and principles. But this requirement harbors some generally unrecognized implications. For if our subjective principles of agency are supposed to prescribe action, then we should be launched on an endless series of choices and prescriptions whenever we attempt to act. Alternatively, our principles might merely express our intentions, or they might otherwise *describe* our actions. But then either the generality of those principles will be threatened, so they cannot function in agency the way formalists have typically supposed, or our autonomy is severely restricted. The following two suggestions for the form and function of maxim principles seem at least consistent with the broader concept of principled free agency, and thus seem like simple solutions to the maxims problem: maxims might be understood to operate in agency more like free agents' intentions in acting *(intentional maxims)*, or more like free agents' dispositions to act *(dispositional maxims)*. I shall explore these alternatives in turn.

Intentional maxims may be a suitable form for the principles of free agency since they are conceived primarily as specific statements of

---

preference principle might be something like: 'Choose to act on your reason for *A-ing* when facing a choice between reasons for *A-ing* and for *B-ing*'. The advantage of this conception of preference principles is that when so conceived they seem not to abrogate prima facie reasons for action, but merely to dictate how to choose between them. Still, in this case the choice between conflicting prima facie reasons, dictated by the higher principle, remains a free choice only in the Aristotelian sense. So the agent's *A-ing*, in the end, should still be seen as a logically necessary consequence of her internalized set of principles, including the preference principle. In the Kantian framework, the preference principle should also be internalized or adopted by a free choice, as with any other maxim

[24] Nagel reveals an awareness of this: "It may be that [prima facie reasons] can be dispensed with in a complete system of practical reason, which will perhaps allow the direct derivation of determinate conclusions about what to do, from the total system of general reasons plus further information" (*op. cit.*, p. 50).

intention. In this form, however, maxims can stand only in one-one relations with actions. Every action will have its unique intentional maxim, implicitly chosen by its agent in the process of forming the intention to act. Indeed, to adopt a maxim of this type can be identified with forming a specific intention to act. These maxims may also be understood as singular reasons for action, if an intention can be a reason. But in that case they are not literally principles of action, as observed previously, and no doubt many formalists will reject the notion of intentional maxims on this account. These maxims might still be regarded somehow as "hypothetical generalities," however. They might at least be *imagined* as principles of conduct: as if an agent acting on a singular, intentional maxim would always act as it indicates in relevant circumstances; or as if everyone would. Generalizing intentional maxims hypothetically in this way might be a useful heuristic for deliberation or moral evaluation. But because their generality is only hypothetical, intentional maxims should not be understood to function in agency in ways that maxims are often thought to function: as personal policies, or as ongoing commitments to courses of conduct. Despite these limitations of intentional maxims, agents supposed to act only on "principles" of this type should at least enjoy complete autonomy in selecting or adopting their maxims of action; for they can never actually determine or commit themselves to act more than once on any of their freely chosen maxims.

Dispositional maxims, on the other hand, may be the only alternative form for the subjective principles of free agency. In contrast to their intentional counterparts, these maxims can be conceived as freely adopted dispositions of character, standing always in one-many relations with actions. Agents who choose by freely adopting dispositional maxim principles seem literally to dispose or determine themselves to the actions indicated in their adopted principles in all relevant circumstances. And while the adoption of a dispositional maxim can be construed as an exercise of autonomy, actions on these maxims can be seen only as necessary consequences of their agents' having freely chosen or adopted them as their principles. But even if choosing to act on an adopted dispositional principle should never be possible, actions on dispositional maxims might be said to be freely chosen at least by courtesy, in virtue of their agents' free choice to adopt the principle.

It may be worth considering whether some middle way can be found between agency principled by singular, intentional maxims, and a life career determined by permanent dispositional principles. A plausible idea is that any free agent's adopted principles should remain capable of being subsequently overthrown. It might always be possible to segue into more refined maxims through free rational reflection. But recall first that freedom is introduced in the formalist model of agency through autonomous, principle-adopting choices: 'Choosing is adopting a maxim'. We cannot therefore *choose* to revise our adopted general principles, for reasons already discussed. Yet unchosen principle revisions might be thought compatible with formalist conceptions of agency. That is, it might be supposed that we still act freely on principles we acquire not by choice, but by rational consideration of reasons. If so, however, then an explanation is required why chosen principles, or *autonomy*, should be necessary for free agency in the first place. We are referred back to the Aristotelian model of agency, where autonomously chosen principles are unknown, and where 'Choosing is being determined to action by principles'.

Second, it should also be observed that anyone fully informed about the revisability of general principles, and conscientious, would never segue into a maxim of the form 'Always to *A* in circumstance *F*'. She would know that expecting herself always to *A* in *F* would be irrational, given the indefinite revisability of her commitments to principles. She might accordingly be satisfied with maxims formulated more like, 'To *A* in *F*, until I change my mind'. But it would probably not be difficult for her to see that considering the nature of her agency, she should never think herself bound to any principle of *A-ing* in *F* beyond the episodes of *F* in which the principle seemed acceptable to her. She would recognize that it could be *her* principle only for the episodes she chooses to act on it. So we are referred, finally, to the model of free agency ruled by *intentional maxims*, where so-called "principles" of agency operate always on a case-by-case basis.

An adequate solution to the maxims problem requires nothing more than conceiving of the form and function of maxim principles in a way that is consistent with a simple requirement on principled free agency: that free choices be conceptually connected with operations on principles. Only intentional or dispositional forms of maxim principles seem suited to this purpose. So in the end, it must be left to formalist theorists to determine whether the subjective principles guiding our free agency are best regarded as functioning singularly and intentionally, or generally and dispositionally. Those developing the Kantian tradition have found it advantageous to see most maxims, if not all, as general or dispositional; though few seem to have noticed the restrictions that general maxim principles must impose on personal autonomy.

RICHARD MCCARTY

East Carolina University

BOOK REVIEWS

Persons and Bodies: A Constitution View. LYNNE RUDDER BAKER. New York: Cambridge University Press, 2000. xii + 233 p. Cloth $52.95, paper $18.95.

John Locke's view that continuants are numerically distinct from their constituting hunks of matter is popular enough to be called the "standard account."[1] It was given its definitive contemporary statement by David Wiggins in *Sameness and Substance*,[2] and has been defended by many since. Lynne Rudder Baker's interesting book contributes new arguments for this view, a new definition of 'constitution', and a sustained application to persons and human animals. Much of what she says develops this view in new and important ways. But in some cases, she does not advance the position; and in others, she takes steps backward.

According to Baker, a person is numerically distinct from her constituting animal. One of Baker's leading arguments is surprisingly unconvincing. Persons differ in important ways from nonhuman animals. Only persons are moral agents, modify their goals, have wars, culture, and so on. If persons were identical to animals—if we were "nothing but animals," as she puts it—then the manifest discontinuity between humans and nonhuman animals would be located "within the domain of biology." "But from a biological point of view, human animals...are biologically continuous with non-human animals" (17). The argument fails: Why should identifying persons with animals preclude saying that these particular animals have radically distinctive features that are of little interest to biologists?

The traditional case for nonidentity (which Baker accepts) is more powerful: a person and her constituting animal differ by having different persistence conditions. If my memories were transferred to a new body and my old body destroyed, I the person might survive, but the human animal who constituted me would perish. Therefore, *before* the transfer, I and the animal that constituted me would be numerically distinct but extremely similar things located in exactly the same place.

---

[1] Michael Burke, "Copper Statues and Pieces of Copper: A Challenge to the Standard Account," *Analysis*, LII (1992): 12-17.

[2] Cambridge: Harvard, 1980.

This consequence—the central thesis of the Wiggins view—is surprising: so surprising that some reject the Wiggins view on that basis. The usual response, that the consequence is unremarkable because the animal *constitutes* the person, only invites the question: What is constitution?

Baker's definition, greatly simplified, is this: x constitutes y if and only if (1) x and y are spatially coincident, and (2) necessarily, anything of x's sort is spatially coincident with something of y's sort (42-43). But constitution, thus understood, cannot explain away the oddness of spatial coincidence, since spatial coincidence is *built into* the definition. We all know Wigginsians think that certain objects (bodies, animals, lumps of clay, and so on) are, when in appropriate circumstances, necessarily co-located with distinct things; the question is how this can be. Labeling the relation of necessitated co-location 'constitution' is no answer. This issue is obscured by Baker's tendentious descriptions of constitution:

> ". .constitution is...a *unity* relation; it is *not mere spatial coincidence*..." (46).
> "...when x constitutes y, there is a *unitary* thing—y, as constituted by x" (46).
> "As long as x constitutes y, x has no *independent* existence" (46).
> "Constitution is as close to identity as a relation can get without being identity" (55).
> "...it is not as if there were two *separate* things—my body and myself. There is a *single* constituted thing—me..." (114).

The italicized phrases (my emphasis) misleadingly portray constitution as an intimate "identity-like" relation. But a quick look back to Baker's definition shows that the suggestive language is unjustified. Stripped of the veneer, constitution is simply necessitated co-location.

Baker's definition faces other problems. Imagine a strange possible world in which the laws of nature do not prohibit two things being in the same place at the same time. In this world, things on a collision course pass right through each other. In such a world, if some object a constitutes another object b, Baker's definition implies that a will also constitute any third object which is the same sort of thing as b and which happens to share spatial location with a, even if it shares no parts in common with a. This is completely contrary to the intuitive picture of constitution. A related problem emerges in worlds with things with no spatial location. The details depend on how exactly "spatial coincidence" is construed. If nonspatial things never spatially coincide, then we get the result that no nonspatial thing could ever constitute another. This is an unwelcome result: assuming we can make sense of nonspatial objects being made up of some sort of nonspatial "stuff," the standard arguments for the distinctness of a continuant and its constituting matter would apply just as well to nonspatial as to spatial objects. On the other hand, if nonspatial things always spatially coincide then any time nonspatial *a* constitutes nonspatial *b*, *a* will also constitute every other nonspatial thing of *b*'s sort.

Most Wigginsians construe constitution as requiring some sort of part-sharing, and with good reason: the counterexamples of the previous paragraph are avoided, and a more satisfying explanation of co-location is possible. So why does Baker not follow this tradition? The reason she gives is that she rejects standard mereology (179-85)—for example, the claim that, if x and y have the same parts, then $x=y$. But rejecting certain theoretical claims about the part-whole relation is no reason to refrain from speaking about parthood altogether!

Baker apparently disagrees with most Wigginsians over how many properties are shared by constitutionally related things (57). For Wiggins, when a lump of clay constitutes a statue, only the lump has the property "being a lump," and only the statue has "being a statue." For Baker, both the lump and the statue have both of these properties. The lump, however, has "being a statue" "derivatively" (46-58). At first glance, this disagreement is superficial. Wigginsians admit an "'is' of constitution," on which the statue "is" a lump in virtue of being constituted by a lump. Read Baker's 'is *F* independently' and 'is *F* derivatively' for Wiggins's 'is *F* predicatively' and 'is *F* constitutively', and the difference might seem to disappear. But Baker makes significant progress on this topic by giving an interesting account of *which* properties are instantiated derivatively and which properties are instantiated independently.

Baker argues that the nature, identity, and essential properties of a thing may be determined by relational features (24, 39, 44). When an artistic community treats a lump of clay as a work of art, there comes to exist an object that is essentially a statue, and so is essentially suitably related to an artistic community. Baker suggests these claims are radical, but I doubt it. Many claim that statues are essentially statues, and I suspect most would agree with Baker that statuehood is extrinsic. Moreover, "orthodox" views about essential properties include origin essentialism and the necessity of distinctness, each of which counts some extrinsic properties as essential.

Baker's discussion of statues brings out one of the more implausible features of her ontology: what there is seems to depend on human interests. Everyone agrees humans can *select* certain objects from a pre-existing stock for their attention. For example, on the four-dimensional ontology I favor, a multitude of space-time worms exist regardless of human activity. Some fall under ordinary predicates: 'statue', 'lump', 'person', 'animal', and so on. But consider odd predicates we do *not* employ—for example, Eli Hirsch's[3] 'incar', which applies only to sums of stages of cars that are contained within a garage. The four-dimensionalist regards these odd objects as just as real as statues and persons; humans simply ignore them. But for Baker, reality includes only the statues and lumps. (If Baker admits incars, and further admits objects corresponding to *all* possible transtemporal tracing concepts, it can be argued that this would amount to admitting temporal stages after all, which she resists.) It is tempting to conclude that she thinks we create the world, that these continuants exist because we have concepts for them. Or is it a cosmic accident that reality just happens to contain objects whose histories match our ordinary concepts for tracing?[4]

Finally, Baker gives her own answer to several outstanding problems with the Wiggins account, including the following. Imagine a lump of clay created already in statue form, which is subsequently squashed. According to Baker, the statue is destroyed while the lump of clay lives on. Therefore there were, all along, a statue and a numerically distinct lump of clay, differing only in historical properties. Critics ask how they can differ in this way when they share the same parts, have the same intrinsic properties, and so on. What *grounds* the difference? Baker's initial answer is that they have different essential properties (169-71). But what grounds *this* difference? Baker replies that her theory is consistent with global supervenience of essential properties on nonmodal properties, since the critics have not shown she is committed to a pair of worlds alike in nonmodal properties but differing modally. This response is promising, I think, but more needs to be said. As I explain elsewhere,[5] the viability of this response depends on distinguishing two often-conflated ways of formulating global supervenience.

Baker's book contains much more than what I have been able to mention here: rich discussions of self-consciousness, personal identity, the importance of persons, and more. It should be read by anyone interested in the metaphysics of persistence and material constitution.

THEODORE SIDER

Syracuse University

---

[3] *The Concept of Identity* (New York: Oxford, 1982), p. 32.

[4] I discuss this argument in more detail in *Four-Dimensionalism: An Ontology of Persistence and Time* (New York: Oxford, 2001), chapter 5, section 3

[5] "Global Supervenience and Identity across Times and Worlds," *Philosophy and Phenomenological Research*, LIX (1999): 913-37.

Lectures on the History of Moral Philosophy. JOHN RAWLS. Cambridge: Harvard University Press, 2000. xxii + 371 p. Cloth $47.50, paper $20.95.

During the years John Rawls was writing his monumental works in political philosophy, *A Theory of Justice*[1] and *Political Liberalism*,[2] he regularly taught an undergraduate course at Harvard University on the history of ethics. This began as a survey of various figures, including Aristotle, Immanuel Kant, and J. S. Mill. In the mid-1970s, however, Rawls changed the course's focus to concentrate primarily on Kant. In her Foreword, Barbara Herman notes that this coincided with an important period in the development of Rawls's own systematic thinking, the work leading up to the 1980 Dewey Lectures: "Kantian Constructivism in Moral Theory" (xiii).[3]

The Dewey Lectures are something of a halfway house in Rawls's thought. On the one hand, they are a work in "moral theory" that outlines a fundamental approach, "Kantian constructivism," that is to be considered alongside other more familiar "traditional moral conceptions, such as utilitarianism, perfectionism, and intuitionism" (*ibid.*, p. 515). On the other hand, Rawls gestures there toward the project that will occupy his published writings from the early 1980s on, "political liberalism," and take him away from public engagement with fundamental moral philosophy. "The aim of political philosophy, when it presents itself in the public culture of a democratic society," he there writes, is to articulate common premises and conceptions that are deeply embedded in a society's public tradition and culture (*ibid.*, p. 518). Although constructivism survives in a form in political liberalism, it is not as a fundamental or "comprehensive doctrine," since these are to be bracketed in the aim of gaining the consensus of reasonable people who, in a pluralistic society, will differ in comprehensive view.

For readers of *A Theory of Justice* who found in its "Kantian interpretation" an exciting new approach to moral philosophy and who welcomed the Dewey Lectures' elaboration of this approach as "Kantian constructivism," Rawls's move away from "comprehensive theory" was not without regret. It was not that political liberalism's approach was unwelcome as political philosophy. The idea that respect for equals within a free society requires a public defense of political principles which is independent of controversial comprehensive pre-

---

[1] Cambridge: Harvard, 1971; revised edition, 1999.

[2] New York: Columbia, 1993; expanded paperback edition, 1996.

[3] This JOURNAL, LXXVII, 9 (September 1980), 515-72

0022-362X/02/9901/49–53

mises can itself be grounded in a Kantian constructivist comprehensive doctrine. But it did mean that Rawls would not be working out such a doctrine in the public sphere of academic moral philosophy.

Nevertheless, at the very time that Rawls's published writing turned toward political liberalism and away from fundamental moral theory, he was achieving in the lectures of his undergraduate ethics course an ever deeper and more refined understanding of constructivism as he found it worked out in Kant. Herman tells us that, in addition to the major revision of 1979, which led to the focus on Kant in the first place, Rawls significantly revised his lectures again in 1987 and 1991 (in the heart of the period between "Justice as Fairness: Political, not Metaphysical" (1985)[4] and *Political Liberalism* (1993)).

Rawls lectured from extensive notes, which he made available to his students. *Lectures on the History of Moral Philosophy* consists of the version of the notes Rawls prepared for the course the last time he gave it in 1991. These are not, however, notes in any ordinary sense. Although their voice has a directness and warmth which is more characteristic of the spoken word, they comprise a genuine book—in my view, one of the most important and distinctive books on the history of ethics ever to have been written.

As befits his legendary modesty, Rawls never intended these notes to be published. Also characteristically, it took an argument from fairness to convince him to do so. Versions of the notes had circulated for years, and Rawls was persuaded that it was unfair that only those with special connections should have access to them.

Some of the outlines of Rawls's interpretation of Kant have previously appeared.[5] But *Lectures* provides a comprehensive treatment—over one-hundred-eighty pages of a marvelously systematic and textually detailed discussion that situates Kant's ethics within his whole philosophy. It combines philosophical interest, insight, systematicity, and textual sensitivity better than any interpretation of Kant's ethics of which I am aware.

Lectures restricts itself to the history of "modern" moral philosophy. The book begins with a discussion of Henry Sidgwick's distinction between ancient and modern ethics. Whereas ethical writers from the seventeenth century on tend to use "quasi-jural notions" like "obligation," "authority," and "imperatives" or "dictates of reason," the ancient Greeks saw ethics as concerning the "attractive," "an ideal good to be pursued" (1). For the ancients, ethics was about "the most rational way to true happiness." For the moderns, it involved the "authoritative prescriptions of right reason, and the rights, duties, and obligations to which these prescriptions of reason give rise" (2). Modern moral philosophers, in other words, were defending a distinctive ethical conception, morality, with its distinctive concept of obligation, and attempting to ground it in the "authority" of reason.

---

[4] *Philosophy and Public Affairs*, XIV (2001): 129-53.

[5] "Themes in Kant's Moral Philosophy," in Eckart Förster, ed , *Kant's Transcendental Deductions* (Stanford: University Press, 1989), pp. 81-113.

Rawls has a fascinating discussion of this difference in fundamental outlook against the differing religio-cultural backgrounds of early modern Europe and ancient Greece. For the Greeks, religion went no further than civic ritual, whereas Christianity involved a salvationist doctrine that claimed absolute authority. The Reformation, moreover, brought the appearance of *rival* authoritative, salvationist doctrines and institutions. Morality, as the moderns came to conceive it in this context, had to involve a law that is binding on all, regardless of these differences in religious doctrine. Rawls is surely right to stress the need to understand this religio-cultural background, no less than the emergence of empirical science and the modern state, if we are to grasp what most seventeenth- and eighteenth-century moral philosophers were up to and, I would argue, its continuing relevance to us today.

After the Introduction, *Lectures* has five chapters on David Hume, two on G. W. Leibniz, ten on Kant, and two on G. W. F. Hegel. The selection of Hume, Leibniz, and Hegel as supporting cast is clearly dictated by the choice of star. Hume's naturalized epistemology and psychologized moral philosophy (including his "fideism of nature") provide a foil for Kant—a picture of our ethical nature without autonomy or genuinely practical reason. Leibniz presents interesting similarities and contrasts. His "philosophy as apology" anticipates Kant's "philosophy as defense," and, also like Kant, he focuses on freedom. But Leibniz is a compatibilist, which Kant thinks permits only the freedom of a "turnspit," and he is a metaphysical perfectionist, which, by Kant's lights, is a form of heteronomy. Finally, Hegel's critique of Kant helps Rawls to bring out certain misconceptions about Kant as well as a way of seeing him as an alternative to both the atomistic individualism Hegel rightly rejects and to Hegel's view of the state as "spiritual substance."

The main idea of a constructivist interpretation is that "practical reason constructs for the will its own object out of itself and does not rely on a prior and antecedent order of values" (230). It follows that the categorical imperative must be viewed as a procedure of construction, the "CI-Procedure," as Rawls calls it. This is clear enough with the first, "universal law," and third, "autonomy," formulations. But neither can the second formulation's injunction to treat humanity as an end in itself be regarded as an independent substantive requirement as, for example, Allen Wood[6] would seek to interpret it. Rather, it is a way of stressing and making more vivid a procedural requirement that is implicit already, namely, that we must "conduct ourselves in ways that are publicly justifiable to [others'] and our common human reason" and offer "such justifications as the occasion demands" (192). Interpreting the various formulations of the categorical imperative as different ways of expressing the requirement to act only on principles that anyone (oneself or others) could will all to act on (that is, will from a common perspective that rules out individual differences) has the great virtue of providing the unification Kant desires without denigrating any of the formulations, as Wood does to the first.

The other virtue this interpretation has is that it connects Kant's normative ethical views intelligibly to his metaethics, specifically, to his conception of the "moral law as a law of freedom." Autonomy, the will's property of being "a law *to* itself," entails not just that there are norms that apply to any and all rational wills, but that there are laws that rational wills necessarily apply or give *to themselves* in the free exercise of their practical reason. It is hard to see how else to understand this idea but procedurally.

But whence freedom? Even if we know that free agents are bound by the categorical imperative, what entitles us to the supposition we are free? Kant shifts from arguing in the *Groundwork* that we cannot help but make this supposition when we act (Ak. 448) to the *Critique of Practical Reason*'s difficult doctrine of the "fact of reason." There the idea is not that we have some independent access to the justified supposition that we are free from which we might deduce the bindingness of moral law. Rather, we start from the idea that we are morally bound, as Kant holds all human beings believe, and see that, because we are, we must be free.

As Rawls reads him, Kant's argument is coherentist at this point. The "fact of reason" cannot be the question-begging simultaneous assertion of moral obligation and freedom. Rather, the supposition of transcendental freedom is necessary for the demands of both theoretical and practical reason (267-68). We should agree with Rawls that this is the best interpretation of Kant's doctrine, I think, but perhaps be less sanguine that the argument can succeed in this form. The problem is that the autonomy of Kantian practical reason seems to go beyond the spontaneity that theoretical reason requires. If theoretical reason's has "as its sole object the understanding and its effective application" (*CPR*, B672), then the spontaneity it needs would not include the capacity to set its own ends and be a law "to itself."[7]

---

[6] *Kant's Ethical Thought* (New York: Cambridge, 1999), pp. 76-155.

In her Foreword, Herman quotes extensively from remarks in which Rawls describes his historical method, notably his insistence, first, always to consider a thinker's views in their strongest form, but only as supported by text, never seeking improvement, and, second, to consider issues as they arose for the thinker, in his own intellectual context, rather than in terms that might help contribute to progress in ours. Both aims are well justified and amply on display in *Lectures*. Nonetheless, it seems overly restrictive to hold that we never profitably study great thinkers like Kant "in the hope of finding some philosophical argument" or idea that might be "directly useful for our present-day philosophical questions" (329). Indeed, some of the depth and plausibility of Rawls's own approach in political philosophy, even political liberalism, seems to me to derive from its relation to fundamental philosophical ideas he finds in Kant.

STEPHEN DARWALL

University of Michigan/Ann Arbor

---

[7] A way of putting the point that is not particularly Kantian is that reasoning about what to believe is governed by a substantive, nonprocedural norm of accurate representation or truth. That nothing analogous holds with practical reason is a reflection, I believe, of the autonomy of the will. I discuss this disanalogy in "Because I Want It," *Social Philosophy and Policy*, XVIII (2001): 129-53.

MEMO TO AUTHORS

The *Journal of Philosophy* is published monthly, in issues that (ordinarily) contain only 56 pages of copy; this limits the length of papers the editors can normally publish to no more than 7,500 words, or eighteen *Journal* pages. A more precise way of determining length is by characters: 2,850 characters per *Journal* page, or 51,300 characters per *Journal* article of 18 pages.

The editors request that *three copies* of all manuscripts submitted be prepared according to the following guidelines:

(1) All material, including the quotations and footnotes, must be typed, double-spaced, and one-sided, with ample margins left and right, top and bottom.
(2) No bibliographies will be published; bibliographical material should be put into the footnotes. No appendices will be published either.
(3) Footnotes should be complete, according to *Journal* style (see any issue). Journal citations should include: name of journal in full, volume number, issue number, month or season, year, and pages. Book references should include: U.S. city of publication, publisher, and date. Footnotes should be gathered at the end of the article
(4) Quotations should be checked by the author; capitalization, italics, spelling, and the like should match the original.
(5) If there are diagrams, they should be provided camera-ready by the author. If there are special symbols, they should be widely spaced and clearly marked. Standard logic symbols are available, and letters of the Greek and script alphabets.
(6) No boldface or underlining; please limit use of italics to technical terms (when they are introduced) and to variables.
(7) No contractions.

The editors are willing to read and evaluate a manuscript that does not meet these guidelines, provided the author agrees to meet them before submitting a final manuscript. The editors may ask for revision of excessively long manuscripts, and may require reformatting of accepted manuscripts. They also request an electronic version with the final hard copy.

The *Journal of Philosophy* does not publish papers that are primarily historical or expository; the editors urge that such manuscripts be submitted elsewhere. While we do not publish manuscripts that mainly comment on articles published elsewhere, we do have a "Comments and Criticism" section in which short responses to *Journal* articles may be published.

Manuscripts may be sent to: Michael Kelly, Managing Editor, *Journal of Philosophy*, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, New York, NY 10027.

You may also visit our website: www.journalofphilosophy.org

[illegible]

INVARIANCES

The Structure of the Objective World

Robert Nozick

In this pathbreaking book, the eminent philosopher Robert Nozick rethinks and transforms the concepts of truth, objectivity, necessity, contingency, consciousness, and ethics. Using an original method, he presents bold new philosophical theories that take account of scientific advances in physics, evolutionary biology, economics, and cognitive neuroscience, and casts current cultural controversies in a wholly new light. Nozick brings together the book's novel theories to show the extent to which there are objective ethical truths.

Belknap Press • cloth

HARVARD
UNIVERSITY
PRESS
800 448 2242
www.hup.harvard.edu

Philosophy comes alive at

Prometheus Books

IS NATURE SUPERNATURAL?

A Philosophical Exploration of Science and Nature

Simon L. Altmann

Has mathematics and physics discovered a new supernatural world, or is this mental cosmos simply an outgrowth of natural evolutionary processes? This is the crucial philosophical issue that Altmann elucidates in this engaging and lucid exposition.

580 pp • ISBN 1-57392-916-6 • HC $39

NATURALISM AND RELIGION

Kai Nielsen

Nielsen makes a strong argument for the cogency and viability of naturalism as the best way of making sense of our lives and our world.

512 pp • ISBN 1-57392-853-4 • HC $40

THE VANQUISHED GODS

Science, Religion, and the Nature of Belief

Richard H. Schlagel

Reviewing both the history of science and the history of Judaism and Christianity as uncovered by modern scholarship, Schlagel comes to the conclusion that the religious viewpoint has been rendered obsolete by scientific advances.

350 pp • ISBN 1-57392-898-4 • HC $36

REINCARNATION

A Critical Examination

REVISED EDITION

Paul Edwards

"It is a breath of fresh air to have at long last a systematic examination of these strange Eastern beliefs. No one could have done the job better than Paul Edwards."

—Martin Gardner

This is the first comprehensive and systematic evaluation of reincarnation and Karma in any language. Edwards' wit and clarity make this a fascinating, accessible, and enjoyable work.

315 pp • ISBN 1-57392-921-2 • PB $23

For exam copies contact editorial@prometheusbooks.com

Call toll free 800-421-0351 • FAX 716-691-0137 • www.prometheusbooks.com

crítica

REVISTA HISPANOAMERICANA DE FILOSOFÍA

Vol. 33 / No 98 / Agosto 2001 / ISSN 0011-1503

Artículos

JESÚS VEGA ENCABO, *Reglas, medios, habilidades. Debates en torno al análisis de "S sabe cómo hacer X"*

GIOVANNA HENDEL, *Realization*

VALERIANO IRANZO, *Bad Lots, Good Explanations*

ANDREW SNEDDON, *What's Wrong with Selling Yourself into Slavery? Paternalism and Deep Autonomy*

Notas bibliográficas

ANA ROSA PÉREZ RANSANZ, *Kuhn y el cambio científico* [León Olivé]

Publicaciones recientes

CRÍTICA, *Revista Hispanoamericana de Filosofía* is published in April, August, and December by the Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México. All correspondence should be addressed to CRÍTICA, Apartado 70-447, Coyoacán, 04510, D.F. México. e-mail: critica@filosoficas.unam.mx

Reenchantment without Supernaturalism

A Process Philosophy of Religion

David Ray Griffin

"David Ray Griffin . . . provides his readers with a full-scale process philosophy of religion. Writing from his own perspective, he discusses in a clear and knowledgeable manner all of the fundamental topics . . . while engaging in conversations with many leading analytic philosophers of religion." —Eugene Thomas Long, University of South Carolina. Cornell Studies in the Philosophy of Religion. $55.00 cloth, $24.95 paper

Pluralism in Philosophy

Changing the Subject

John Kekes

This original and ambitious book aims to change how we think about good lives. The perennial debates about good lives—the disagreements caused by conflicts between scientific, religious, moral, historical, aesthetic, and subjective modes of reflection—typically end in an impasse. Kekes defends a pluralistic alternative to absolutism and relativism that will, he holds, take philosophy in a new and more productive direction. $32.50

Soul, Body, and Survival

Essays on the Metaphysics of Human Persons

EDITED BY
Kevin Corcoran

Kevin Corcoran's collection brings together cutting-edge research on the metaphysics of human nature and the possibility of post-mortem survival. It includes chapters from those who embrace traditional soul-body dualism, those who assert person-body identity, and those who propose entirely new views. $45.00 cloth, $19.95 paper

Life's Form

Late Aristotelian Conceptions of the Soul

Dennis Des Chene

"Dennis Des Chene has trawled the natural-philosophy and metaphysics literature on matter theory and the relation between the mind and the body, offering a thorough and sophisticated discussion of individuation and differentiation of plants, animals, and human beings in late scholastic thought." —Stephen Gaukroger, University of Sydney. $45.00

Spirits and Clocks

Machine and Organism in Descartes

Dennis Des Chene

"An enormously impressive accomplishment. In this book, Dennis Des Chene brings together the history of biology, the history of mechanics and late scholastic philosophy to illuminate the way in which Descartes transformed our understanding of living creatures." —Daniel Garber, University of Chicago. 20 illustrations. $39.95

New in Paper—

Physiologia

Natural Philosophy in Late Aristotelian and Cartesian Thought

Dennis Des Chene

Winner of the Morris D. Forkosch Prize from the Journal of the History of Ideas. "For the growing number of scholars with a serious interest in late scholasticism and its relationship to early modern philosophy, the book should be very stimulating and a rich source of information." —*Philosophical Review*. $19.95

At bookstores, or call (607) 277-2211 • www.cornellpress.cornell.edu

The Metaphysics of Beauty

Nick Zangwill

In this book, Nick Zangwill investigates the nature of beauty as we conceive it, and as it is in itself. Zangwill argues that it is essential to beauty that it depends on the ordinary features of things. He uses this principle to defend the notion of the aesthetic, to call for a version of aesthetic formalism, and to reconsider the reality of beauty. $39.95

New in Paper—

The Engine of Visualization

Thinking through Photography

Patrick Maynard

"For Patrick Maynard photography is in the first instance a technology for marking surfaces, and from this insight flows the most original sustained analysis and argument yet produced in the history of thinking about the camera and its work. At once analytical and historical, scientific and aesthetic . . . [this book] spells out the simplest general principles necessary for a clear understanding of what photography—also technology, image, and art—is and does." —Alan Trachtenberg, Yale University.

33 illustrations $19.95

The Practice of Persuasion

Paradox and Power in Art History

Keith Moxey

Offering a series of meditations on the discipline of art history in the context of contemporary critical theory, Moxey addresses such central issues as the status of the canon, the nature of aesthetic value, and the character of historical knowledge. 7 illus. $39.95 cloth, $15.95 paper

Performing Live

Aesthetic Alternatives for the Ends of Art

Richard Shusterman

Ranging from rap, techno, and country music to cinema, cyberspace and urban design, Shusterman develops his radical theory of "somaesthetics," charting the complex network of bodily arts so prominent in contemporary life and self-styling. By blending concrete aesthetic analysis with insightful social critique, Shusterman provides a rich menu and critical guide for today's pursuit of the art of living. $39.95 cloth, $17 95 paper

The Life of a Style

Beginnings and Endings in the Narrative History of Art

Jonathan Gilmore

"An exciting cross-fertilization of analytic philosophy and current art history, this book applies theories of history and narrative drawn from the philosophy of biology, personal identity, and literature and offers well-informed and sensitive interpretations of several exemplary artistic traditions." —Whitney Davis, Northwestern University

31 photos. $35.00

Cornell University Press

Sage House 512 East State Street, Ithaca, New York 14850

INFERENCE FROM SIGNS

Ancient Debates about the Nature of Evidence
JAMES ALLEN

"Allen's brilliant investigation is a masterpiece of historical as well as philosophical scholarship adds a new dimension to our understanding of Hellenistic epistemology...will quickly become the authoritative work on this important chapter in the histories of science and philosophy'"
—Gisela Striker, *Harvard University*

2001 296 pp $45 00

DETERMINISM AND FREEDOM IN STOIC PHILOSOPHY

SUSANNE BOBZIEN

"[Bobzien] demonstrates a scholarly command of the ancient texts and the contemporary secondary literature that places her as one of the most knowledgeable philosophers working in the history of ancient philosophy today [This book] is likely to be a dominate work in the field for quite some time .."—*The Review of Metaphysics*

1999 (paper September 2001) 456 pp paper $26 00, cloth $85 00

THE ETHICS OF KILLING

Problems at the Margins of Life
JEFF MCMAHAN

"[This] is applied ethics at its best From now on, anyone who is serious about getting to the bottom of issues like abortion, infanticide, brain death, euthanasia and the killing of nonhuman animals will have to take account of the novel and ingenious theory presented in Jeff McMahan's lucidly-written, rigorously-argued book"
—Peter Singer, *Princeton University*
(Oxford Ethics Series)

December 2001 624 pp. $45 00

THE ANIMAL QUESTION

Why Non-Human Animals Deserve Human Rights
PAOLA CAVALIERI
Translated by CATHERINE WOOLLARD

This fascinating, provocative work advances a radical proposal that we extend basic human rights to the nonhuman animals we currently treat as "things"

"A brilliant, concise statement of the argument for attributing basic rights to animals."—Peter Singer, *Princeton University*

November 2001 208 pp $25 00

PASSIONATE ENGINES

What Emotions Reveal about the Mind and Artificial Intelligence
CRAIG DELANCEY

DeLancey shows that our understanding of emotion provides essential insight on key issues in philosophy of mind and artificial intelligence He offers us a bold new approach to the study of the mind based on the latest scientific research and provides an accessible overview of the science of emotion

November 2001 304 pp. $49 95

COMMON SENSE, REASONING, AND RATIONALITY

Edited by RENEE ELIO

While common sense and rationality often have been viewed as two distinct features in a unified cognitive map, this volume offers novel, even paradoxical views of the relationship Touching on various disciplines, it considers what constitutes human rationality, behavior, and intelligence
(New Directions in Cognitive Science)

December 2001 352 pp., 15 line illus paper $29 95 cloth $70 00

ESSAYS ON PLATO AND ARISTOTLE

J. L ACKRILL

"These seminal papers remain required reading for those working in the field The book serves as a fitting testimonial to the career of one who has had a key role in shaping the study of ancient philosophy as it is pursued today"—*Bryn Mawr Classical Review*

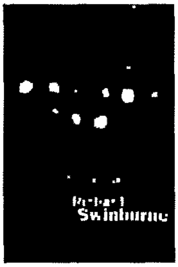
1997 (paper 2001) 264 pp paper $22 95 cloth $60 00

OXFORD

Prices are subject to change and apply only in the US To order, or for more information, please call 1-800-451-7556 In Canada, call 1-800-387-8020
Visit our website at **www.oup.com**

EPISTEMIC JUSTIFICATION

RICHARD SWINBURNE


Richard Swinburne

What makes a belief rational or justified in holding? Swinburne maps various accounts of epistemic justification, arguing that most kinds of justification are worth having because they are indicative of truth However, only justification of internalist kinds can guide a believer's actions

2001 305 pp paper $19 95 cloth $55 00

AUTONOMOUS AGENTS

From Self-Control to Autonomy

ALFRED R. MELE

Mele argues that even an ideally self-controlled person can fall short of personal autonomy and examines what needs to be added to such a person to yield an autonomous agent. " Mele has hit his mark in this well-argued, engaging, and thought-provoking book "—*The Review of Metaphysics*

1995 (paper 2001) 288 pp paper $19 95 cloth $60 00

Chosen as an Outstanding Academic Book of 1997 by Choice

THE LAST WORD

THOMAS NAGEL

" [a] subtle, compact, and forceful book a work of philosophical reflection a significant contribution to the culture wars of our time"—*New York Review of Books*

" . a ringing defense of the rationalist conception of reason a book that should be read and pondered in this golden age of subjectivism"—Colin McGinn, *The New Republic*

1997 (paper 2001) 160 pp. paper $11 95 cloth $22 00

OBJECTS AND PERSONS

TRENTON MERRICKS

There are no statues Or rocks Or chairs. But there are microscopic objects arranged statuewise and rockwise and chairwise Moreover, there are—in addition to microscopic objects arranged humanwise—composite human beings Or so Trenton Merricks argues in this lucid and provocative book

2001 216 pp $35 00

WITTGENSTEIN IN AMERICA

Edited by TIMOTHY G MCCARTHY and SEAN C STIDD

This remarkable collection explores the legacy of Wittgenstein's work in contemporary American philosophy The contributors take a variety of approaches in their discussion of such topics as rule-following, realism in mathematics, and Wittgenstein's distinction between sense and nonsense

2001 240 pp $45 00

GENDER IN THE MIRROR

Cultural Imagery and Women's Agency

DIANA TIETJENS MEYERS

Harmful, culturally prevalent feminine imagery constrains women's self-determination. Meyers proposes alternative imagery and advances an account of feminist discursive politics that takes on the challenge of neutralizing patriarchal imagery

(Studies in Feminist Philosophy)

December 2001 256 pp ; 8 photos paper $19.95 cloth $55 00

CONFUCIUS AND THE ANALECTS

New Essays

BRYAN W VAN NORDEN

This first English volume on the subject employs philosophical, philological, and religious methodologies to address a number of topics (from Confucius and Western "virtue ethics" to his attitude toward women) important to the historical composition of the text of the *Analects*

November 2001 348 pp , 4 line illus paper

Prices are subject to change and apply only in the US To order, or for more information, please call 1-800-451-7556 In Canada, call 1-800-387-8020 Visit our website at **www.oup.com**

OXFORD

NEW PHILOSOPHY FROM CHICAGO

THE WORK OF MOURNING

Jacques Derrida

Edited by Pascale-Anne Brault and Michael Naas

"This astonishing book serves both as an introduction to contemporary French thought in all its depth and breadth and as a meditation on what it means to mourn the passing of friends and colleagues. Anyone who happens to be mortal and who has been forced to bear witness to the mortality of others should read this thoughtful and deeply felt book."—David Farrell Krell, author of *The Good European*

CLOTH $25.00

METAPHYSICAL HORROR

Leszek Kołakowski

Edited and translated by Agnieszka Kołakowska

"[A] philosopher who more convincingly than any other I can think of demonstrates the continuing significance of his vocation in the life of our culture."—Karsten Harries, *New York Times Book Review*

PAPER $16.00

THE SPELL OF LANGUAGE

Poststructuralism and Speculation

Thomas G. Pavel

Translated by Linda Jordan

"One of the most rewarding books I have read. . . So comprehensive is [Pavel's] understanding that this . . is an excellent guide to the historical developments of linguistic science."—*Philosophical Books*

PAPER $16.00

KANT, HERDER, AND THE BIRTH OF ANTHROPOLOGY

John H. Zammito

"This is intellectual history at its best. John Zammito has accomplished two important goals in his new book. By placing Kant in a rich and underappreciated historical context, Zammito strips away the veneer of inevitability surrounding most accounts of Kant's critical project. Second, by taking Herder's philosophy seriously, Zammito places Herder back where he belongs: in the pantheon of the great eighteenth-century thinkers. A major achievement."—Robert E. Norton, University of Notre Dame

PAPER $29.00

ISAIAH BERLIN

A Celebration

Edited by Edna Ullmann-Margalit and Avishai Margalit

"[T]he collection testifies to the learning and profundity of Berlin's thought and, by way both of reminiscence and influence, to the charm and gaiety of its expression."
—Anthony Quinton, *The Times* of London

PAPER $18.00

THE UNIVERSITY OF CHICAGO PRESS
1427 East 60th Street, Chicago, IL 60637 • www.press.uchicago.edu

THE JOURNAL OF PHILOSOPHY

ALIPUR CAMPUS

16/5/02

VOLUME XCIX, NUMBER 2
FEBRUARY 2002

page
55 *Two Ways of Thinking about Fitness and Natural Selection* Mohan Matthen and André Ariew
84 *Causation: Influence versus Sufficiency* Michael McDermott

BOOK REVIEWS
102 Ruth Garrett Millikan: *On Clear and Confused Ideas: An Essay about Substance Concepts* Robert Cummins, Alexa Lee, Martin Roth, David Byrd, and Pierre Poirier

109 NEW BOOKS: TRANSLATIONS

***Published by* The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027 FAX: (212) 932-3721. You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2002

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 2002 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCIX, NO. 2, FEBRUARY 2002

TWO WAYS OF THINKING ABOUT FITNESS AND NATURAL SELECTION*

The concept of fitness is, Philip Kitcher[1] says, "important both to informal presentations of evolutionary theory and to the mathematical formulations of [population genetics]" (*ibid.*, p. 50). He is absolutely right. The difficulty is to harmonize these very different ways of understanding its role. We examine here how fitness and natural selection relate to the other explanatory factors invoked by evolutionary theory. We argue that the "informal presentations" to which Kitcher alludes give an incoherent account of the relation. A more appropriate model is drawn from the statistical framework of population genetics. We argue that this model demands a far-reaching revision of some widely accepted notions of causal relations in evolution.

I. VERNACULAR FITNESS AND PREDICTIVE FITNESS

The two formulations of fitness are well illustrated by expositions given by Kitcher himself. Consider:

> *The principle of variation in fitness*: Organisms differ in ways that affect their competitive abilities. Some organisms have characteristics that better enable them to survive and reproduce than others (*ibid.*, p. 38).

---

* This article was originally presented before the 2001 Congress of the Canadian Philosophical Association, and helpfully commented on there by Wayne Myrvold. We acknowledge helpful discussion with Paul Bartha, Bill Harper, Andrew Irvine, Dick Lewontin, Patrick Maher, Joel Pust, Patrick Rysiew, Larry Shapiro, Elliott Sober, Michael Strevens, and Catherine Wilson.

[1] *Vaulting Ambition: Sociobiology and the Quest for Human Nature* (Cambridge: MIT, 1984) Elliott Sober makes a similar distinction. "Fitness is both an ecological descriptor and a mathematical predictor"—"The Two Faces of Fitness," in R. Singh, D. Paul, C. Krimbas, and J. Beatty, eds., *Thinking about Evolution: Historical, Philosophical and Political Perspectives* (New York: Cambridge, forthcoming). Kitcher and (even more so) Sober imply that the underlying logic of these concepts is somehow the same, though Sober acknowledges persistent difficulties drawing them together.

0022-362X/02/9902/55–83

An organism's ability to "survive and reproduce" arises from its traits. To the extent that relatively "advantageous" traits can be inherited by an organism's descendants, they will be reproduced and retained in the population at a higher rate than less optimal ones. Consequently, these traits will spread faster than less optimal ones. The resulting differential growth of gene frequencies is natural selection. This much is central to common-sense analysis, and for many this notion of an organism's *overall competitive advantage traceable to heritable traits* is at the heart of the theory of natural selection. Recognizing this, we shall call this measure of an organism's selective advantage its *vernacular fitness*. According to one standard way of understanding natural selection, vernacular fitness—or rather the variation thereof—is a *cause* of evolutionary change. Note, however, that, as formulated above, vernacular fitness is merely *comparative*, not *quantitive*, and that principles such as the above afford us no way of predicting or explaining the *magnitude* of evolutionary change. This is why it is appropriate to call it *informal*.

As Kitcher observes, this is not the only context in which the concept of fitness crops up in evolutionary theory. Fitness occurs also in equations of population genetics which predict, with some level of probability, the frequency with which a gene occurs in a population in generation $n+1$ given its frequency in generation n. In population genetics, *predictive fitness* (as we shall call it) is a statistical measure of evolutionary change, the *expected* rate of increase (normalized relative to others) of a gene, a trait, or an organism's representation in future generations, or, on another interpretation, its *propensity* to be represented in future generations, suitably quantified and normalized.[2]

Unlike vernacular fitness, predictive fitness is not a cause of selection, or of evolution for that matter. The expected rate of return on an investment is not a *cause* of its growth—it just *is* its growth, numerically estimated and projected into the future. Likewise, predictive fitness is a measure of evolutionary change not a cause. Nor do these conclusions have to be modified much if we treat these quantities as *propensities*. An investment's propensity to grow is a consequence of all the factors that economists identify as causes of growth—profit, growth prospects, dividends, and so on. From an economist's point of view, explaining an investment's propensity to grow is no different from explaining its expected growth. In exactly the same way, even if predictive fitness is thought of as a propensity, and so a cause of evolutionary change, it is so closely identified with that change as to make no difference from a scientific point of view—explaining one is the same as explaining the other.

---

[2] We do not mean to insist on the expected-value interpretation; there are many ways of defining predictive fitness within population genetics. For a review, see John Endler, *Natural Selection in the Wild* (Princeton: University Press, 1986).

Now, what is the relationship between vernacular fitness and predictive fitness? As we have seen, conventional wisdom holds that vernacular fitness is a cause of evolution. If predictive fitness measures the rate of evolutionary change, then, vernacular fitness might be thought to be a cause of predictive fitness, a part of its biological *why*, much as the shape, composition, and symmetry of a coin are physical causes of its being "fair," this being (like predictive fitness) a statistical characterization of its behavior when tossed. In other words, one is tempted to say that part of the reason a particular gene tends to increase in frequency is that it has a competitive advantage. On the other hand, vernacular fitness is not the *only* thing responsible for evolution. Such other factors as the underlying mechanism of inheritance and drift must be taken into account, as well as various constraints on selection. These supposedly "nonselective" factors would need to be combined with vernacular fitness in any adequate account of the "why" of predictive fitness. How are they to be combined? Any attempt to relate vernacular fitness to predictive fitness, and thus to harmonize Kitcher's "informal" and "mathematical" discourses, needs to come to grips with this question.

One standard approach to this question is to say that the contribution vernacular fitness makes to evolution can be understood by means of an analogy with the standard physical-science conception of multiple forces summed up to produce a resultant total force. This way of addressing the harmonization problem is one of the "two ways of thinking" to which our title alludes. In sections II through IV, we take issue with this analogy. We argue that it makes no sense to *add* natural selection to other evolutionary factors in this way; the combination of factors is better described by the *hierarchical-realization* model presented and discussed in sections V through VIII. We argue further, in section IX, that natural selection is not a process driven by various evolutionary factors taken as forces; rather, it is a statistical "trend" with these factors (vernacular fitness excluded) as predictors.[5] These theses demand a radical revision of received conceptions of causal relations in evolution.

---

[5] The distinction between physical process and statistical trend parallels a distinction between dynamical and statistical analyses in Denis Walsh, Tim Lewens, and Ariew, "The Trials of Life: Natural Selection and Random Drift" (in preparation).

II. THE THEORY OF EVOLUTION AS A THEORY OF FORCES

Elliott Sober[4] is the author of one influential suggestion about how the theory of evolution accommodates natural selection alongside drift, developmental constraints, architectural constraints, and other such determinants of evolutionary change. He proposes that we should think of the theory of evolution as a theory of forces along the lines suggested by Newton's mechanics (*ibid.*, p. 31), and of natural selection (or vernacular fitness) as *one* (but only one) of the forces that may act on a population.

Sober urges that the following structural elements of Newton's theory be incorporated into the theory of evolution.

(1) *A zero-force law.* Such a law equates the steady-state, or no-change condition to no net force. In classical mechanics, a body is not changing if it continues in its state of rest or steady motion. Newton's first law of motion stipulates that a body is in this steady state exactly when no force is impressed upon it. Sober claims (*ibid.*, p. 32) that the corresponding steady state in evolution is specified by the Hardy-Weinberg law of population genetics. This is not exactly right—the Hardy-Weinberg law tells us about genotype frequencies only in *sexually reproducing* populations when there is no change of gene frequencies within them. The most general way of capturing Sober's intentions is to stipulate that the steady state occurs when gene frequencies do not change from one generation to the next. Following the Newtonian analogy, there is no net force acting on a population in this condition. This would be the zero-force law for evolutionary theory.

(2) *Decomposition of force.* Newtonian physics envisages several sources of force—gravitation, electrostatic force, magnetic force, impact, drag, and so on. In classical physics, "source laws," as Sober calls them, describe how these forces emerge, one by one, from various physical situations. The law of gravitation, Coulomb's law, the law of elastic bodies, the laws of aerodynamic resistance, and the like are examples of such source laws. The total force acting on a body (the force that feeds into the consequence laws described under (4), below) can always be decomposed into these components. Sober suggests (*ibid.*, p. 38) that selection, mutation, migration, random genetic drift, linkage, and in-breeding are the individual forces that act on a population and cause it to evolve. Laws governing these are

---

[4] *The Nature of Selection: Evolutionary Theory in Philosophical Focus* (Cambridge: MIT, 1984), chapter 1. Since a large part of this article is devoted to taking issue with Sober, we want emphatically to acknowledge our personal debt to him. We, like many other philosophers, learned much of what we know about natural selection from Sober's classic work. We should also like to acknowledge his patient discussion of the issues here

source laws: for example, the principles of optimality analysis which predict the selective value of various traits and the laws that govern how genes located on the same chromosome assort during recombination. Accordingly, causal responsibility for total evolutionary change can be apportioned among these factors.

(3) *Resolution of forces.* Given two or more forces acting on a body, Newtonian physics uses the parallelogram of forces (that is, vector addition) to compute the total resultant force acting on the body. Evolutionary forces combine in this way, according to Sober. He offers the case of *heterosis* to illustrate the point, contending that here selection is added to the genetic recombination found in sexual reproduction (*ibid.*, p. 47). (A description of this phenomenon, and our own analysis, will be found in section v.)

(4) *A law describing change under resultant force.* Newton's second law of motion ($F = ma$) tells us that the change of a body's momentum is proportionate to the total resultant force impressed on it. Sober refers to this as a "consequence law." The laws of population genetics are supposed to play this role in evolutionary theory.

Now, it is of central importance in Newton's theory that the mathematical device of "resolving" forces by vector addition does not compromise the *separate* operation of "component" forces like gravitation and drag. Newton's second law of motion predicts change given force. It makes no difference, however, whether we first combine forces by vector addition and then use Newton's second law to derive acceleration, or, reversing the order of these operations, first feed the component forces into the second law one by one and then use vector addition to combine the separate acceleration vectors that result. In other words, acceleration under the sum of component forces is just the same as the sum of accelerations due to the component forces taken separately. Resultant force is, in this respect, dispensable. (It is needed, however, to preserve the universality of Newton's second law by positing a force that corresponds to total acceleration.) Physically, the component forces act independently of one another; there is no mechanism that creates a new force out of them.

This independence condition for component causation, which we regard as indispensable to the Newtonian apparatus of resolved force, fails in the evolutionary case. The consequence laws are, as we have seen, the index of total evolutionary change in population genetics. So, if Sober's conception of evolutionary forces is right, one should expect that vernacular fitness and the other causes of evolution add up to a resultant force. If the analogy with Newtonian mechanics is to be maintained, these components should, despite such summation, retain their separate causal influences. But, as we shall argue in the next two sections, *it is incoherent to think of the component factors contrib-*

uting to evolutionary change by separate action. As a consequence, the analogy with Newtonian forces collapses: the first three structural elements mentioned above find no counterpart in evolutionary theory, and the fourth, the notion of a consequence law, makes sense only in terms of probabilistic quantities like predictive fitness, with no role for vernacular fitness.[5]

III. DRIFT AND THE ZERO-FORCE LAW

The trouble starts with the zero-force law. In Newton's theory, inertia—the property of each massive body to resist changes to its "state of rest or uniform motion along a straight line"[6]—ensures that, if there is no force acting on a body, then there is no change in its momentum. But prima facie at least, this is not the case in evolution. For there are chance variations in the genetic composition of a population from generation to generation. These changes can have a dramatic cumulative effect. According to the theory of the "random walk," an entity that is varying under chance influences will drift further and further away from its starting condition, the cumulative deviation being proportionate to elapsed time. It follows that the frequency of a given gene will, with time, drift further and further away from its starting frequency, and given enough time, it will approach a frequency of either zero or one as chance has it. But once it reaches zero or one, the change is (barring regeneration by mutation) irreversible: either the trait has gone extinct or it has reached "fixation." Thus, given a set of alleles, one will eventually go to fixation even without the influence of fitness differences. But this creates a problem for the zero-force law. In order to save it, Sober claims that this kind of change is caused by "drift," which he therefore counts among the "forces" of evolution.

> In evolutionary theory as in Newtonian physics, the principal use of a zero-force law is to discover when evolutionary forces *have* played a role. If genotype frequencies depart from. .equilibrium, some force must have been at work.. . "[S]ampling error" (*random genetic drift*) may lead the frequency of a gene among the zygotes to differ from the frequency of that gene among the gamete pool.[7]

But does it really makes sense to say that drift is a *force* or, more generally, a cause of change that acts independently of selection? Consider this analogy. You toss a coin four times. What would explain the outcome *two heads*? Answer: the physical setup of the coin-tossing trials. What would explain the outcome *four heads*? Answer: the same thing. Although the second result is less probable, the same setup explains both outcomes. The set of possible worlds in which fair coins are tossed four times form what Wesley Salmon[8] calls a "homogeneous reference class"; there is no way to partition such a reference class by an independent relevant factor such that the probability of four heads is greater in one of the cells of the partition.[9] Perhaps one might say that the four-head outcome was less predictable, therefore less well explained within this setup. However that may be, there is nothing available with which to strengthen the explanation of the four-head outcome. In particular, you cannot partition the homogeneous reference class to which these trials belong by *improbability* or *chance*, and so you cannot tag these things as "forces" that occasionally favor the less likely outcome.

---

[5] Michael Strevens forced clarity here, and in the section I discussion of predictive fitness.

[6] *Principia*, Definition III.

[7] *The Nature of Selection*, p. 34.

For closely related reasons, "drift" should not be regarded as a force that can be added to others acting on a population. (In any case, drift is not the sort of thing that can play the role of a force—it does not have predictable and constant direction.) Imagine allelic alternative traits T and T', such that T confers more causal fitness on its bearer. Suppose that T and T' compete in two similar populations of conspecific organisms, subject to the same pressures of selection. In one population, T goes to fixation, in the other, T' does. What explains this difference? The answer, just as in the case of the coin, might well be: *exactly the same thing*, the same positioning of the traits in the adaptive landscape. Maybe the explanation is weaker in the case of T' going to fixation, particularly if the fitness difference is large. But it does not follow that any other explanatory factor is available to help out. So one cannot say that "if genotype frequencies depart from equilibrium, some force must have been at work." Because the causes here are probabilistic, change might have the same cause as equilibrium.

Does this miss the point? Sober says (see above): "In evolutionary theory as in Newtonian physics, the principal use of a zero-force law is to discover when evolutionary forces *have* played a role." And this might be taken to mean that one can retrospectively identify drift in particular evolutionary histories, not simply in stochastic aggregates.

---

[8] *Scientific Explanation and the Causal Structure of the World* (Princeton: University Press, 1984), pp. 36-37.

[9] That is: if C is the condition that defines the homogeneous reference class to which this trial belongs, then the probability of the target explanandum *four heads* is the same (1/16) given C as it is given C *and* any other putative factor, F.

Histories of natural selection consist, after all, of collections of concrete individual events: births, deaths, matings, mutations, and so on. Some of these events are predictable on the basis of advantageous traits (that is, vernacular fitness), others are not. In the first kind of case, we have, as some say, *discriminate* sampling, in the second, *indiscriminate*. And so one might think that one can fix the exact role of vernacular fitness and of drift by looking more minutely at individual events and determining when discriminate sampling has been at work, and when indiscriminate. In this way, one apportions the outcome to fitness and drift, respectively.

Consider, then, a case like this: two organisms, O1 and O2, otherwise very similar, differ in (vernacular) fitness because O1 has better eyesight than O2. Now, contrast the following possible events.

(C1) O2's bad eyesight leads to its falling off a cliff. It dies and O1 survives.

(C2) O1 is killed by a lightning strike—the difference of visual acuity was irrelevant to this event.

There is evolutionary change in both these cases, because both lead to a change in the gene frequency of the population: in (C1), O1's visual acuity genes get a slight boost; in (C2), they suffer a small setback. (C1), however, seems to be a case in which the difference of vernacular fitness (the difference in eyesight) contributed to evolution, and (C2) one in which a chance event thwarted the fitness difference that drives natural selection. So it may be tempting to say that natural selection is the cause of evolution in (C1), and that it consists, over a longer period of time, of "predictable" (or fitness-biased) cases like (C1), but that it excludes anomalous (or fitness-indiscriminate) cases like (C2). Then it might be thought plausible to say that something else—drift? indiscriminate, or neutral, selection?—is operating in (C2).

But this violates sound probabilistic thinking. Probability enters into the picture because the theory of evolution abstracts away from concrete individual events like (C1) and (C2) in order to isolate the causal factors that make a probabilistic *difference* to evolution.[10] Since, ex hypothesi, your chances of being killed by a lightning strike are the same whether you have good eyesight or bad (other factors being held equal), this factor cannot be conjoined with either good or bad eyesight to yield finer predictions. This does not mean that evolutionary theory assigns cases like (C2) to some special category of "uncaused events." Lightning *has* been taken into account when we say (a) that bad eyesight and the like causes earlier death with probability p (less than one), and further (b) that there is *no other factor* that will enable us to refine this estimate.[11] That is, it has been taken into account when we judge that there is no variance among individuals in terms of their vulnerability to lightning strikes.

---

[10] There are several kinds of situations in which the notion of probabilistic causation is invoked. One is in quantum mechanics, where it is claimed that the interaction of certain variables is irreducibly indeterministic. Thus, if *QM* variable X brings about result Y with probability $p\%$ (less than unity), it is unclear whether one should say that X is 100% involved in $p\%$ of the cases, thus bringing about Y, and inert in the rest, or rather that X is $p\%$ involved in 100% of the cases, and brings about Y in $p\%$ of the cases as a result. But let us leave this example aside, along with the complications it brings. Indeterminism is not the issue in the cases that we are

This highlights a confusion of levels which can occur when we speak of probabilistic causes. On the one hand, we say that lightning strikes are not a factor in natural selection. By this, we mean something like the following: lightning strikes do not, over large populations, over the long haul, result in the differential retention of heritable traits. This is what we shall call an *ensemble*-level fact, or *e*-fact: it is concerned with frequencies, probabilities, or propensities which derive from aggregates. On the other hand, we need to acknowledge that *any* death, from whatever cause, results in evolutionary change since it results in a change of gene frequencies. Consequently, we can say that lightning strikes cause evolutionary change in cases like (C2). This is a statement which derives from a causal relation at the level of concrete *individuals*, not ensembles: it is an *i*-fact. The pair of statements seems contradictory only when the difference of levels goes unnoticed.

Now, it is often not sufficiently well understood in discussions of causation that there is no straightforward translation between levels with regard to causal statements. Suppose that good eyesight causes individuals like X to live n years longer with probability m. This statement has something to do with proportions of X-like individuals who live n years longer in various possible worlds. It states an e-fact. Now consider a particular individual X, who has good eyesight and has lived a long life. Here we are considering an i-fact. We might wonder what exactly the e-fact just alluded to tells us about the causation of the i-fact concerning X. How many years did good eyesight contribute to *her* long life, and how much of her life should we attribute to good luck? Generally, this question has no determi-

---

discussing. As with the fair coin, one would not be justified in claiming that the individual events above contained elements of indeterminacy.

[11] We are indebted here to Andrew Irvine and Joel Pust for helpful discussion.

nate answer. When one is talking about *types* of events, it makes sense to quantify the role of chance. For once a reference class has been partitioned in terms of all the factors that make a difference, the residual variation within the cells of the partition—the unassigned variation—is uncaused. It is due to chance if you like. "Unassigned variation" has no i-counterpart, however. Although it is six times more probable that two heads will turn up in a run of four tosses of a coin than that four will, chance does not play any more of a role in a particular run of four heads than in a particular run of two and two. Thus, one cannot in general differentiate between individual events on the basis of how much they are attributable to chance.

For exactly the same reason, one cannot *in general* determine, even retrospectively, whether discriminate or indiscriminate sampling was responsible for an actual sequence of historical events. In an example discussed above, T and T' were traits that suffered opposite fates in two closely comparable populations. We can certainly estimate the probability that T will win or that T' will, counting these as event types. We also know things like this: the chances of the less viable trait going to fixation are greater if the population is smaller—this application of the "law of large numbers" is the stochastic basis for the "founder effect," that is, for novelty arising when small populations are isolated. These statements are based on the assignment of unassigned variation to event types. There is, however, no such thing as unassigned variation with respect to *concrete* sequences of events (barring true indeterminacy). Suppose, then, that one is trying to explain a particular biological characteristic—for instance, the absence of body hair in humans. It would be correct to say that this characteristic probably arose in a small population, where novelty had a better chance of taking hold—*but only if this is meant as a statement about types* to which we revert because we lack certainty about evolutionary history. It would be a mistake, however, to ask how much chance contributed to the actual history of human hairlessness, just as it would be a mistake to try to apportion the role of chance in a particular series of coin tosses.[12] Even if we are able to identify each and every i-event in that history, and determine which involved the (deterministic) operation of factors, like good eyesight, that make an

[12] Alexander Rosenberg—*Instrumental Biology or The Disunity of Science* (Chicago: University Press, 1994)—has recently argued that drift enters into evolutionary theory only because we are ignorant of particular causes. He is right if he is talking about historical reconstructions of evolutionary scenarios. But he is wrong if he is talking about evolutionary *theory*, which abstracts away from individual causes.

e-level difference, it will still not make sense in general to quantify the residual role of chance.[13]

The theory of evolutionary forces has a problem. In the passage quoted above, Sober rightly demands that one ought to be able to tell when a force is acting on a system. He puts the point in epistemological terms, but one can rephrase it ontologically. *Whether there is a force acting on a system should be a matter of fact.* This, however, is precisely the question that may not be well defined in this instance. Suppose that over a period of time a population stays exactly the same, or changes in some determinate way. The proposition that drift was involved to degree p in this history generally has no determinate truth value.

IV. SEPARATING EVOLUTIONARY FORCES

Continuing our examination of evolutionary theory as a theory of forces, we now ask how one should go about decomposing evolutionary force. Here it is useful to introduce another context in which the notion of component force has sometimes been used. Stephen Jay Gould and Richard Lewontin[14] are associated with the claim that to appeal to natural selection alone as an explanation of evolutionary change, and to underestimate the other factors, is a methodological error to which they gave the name "adaptationism." They define this error as the belief that "natural selection [is] so powerful and the constraints upon it so few that the direct production of adaptation through its operation becomes the primary cause of nearly all organic form, function, and behavior" (*ibid.*, p. 256).

For the sake of the present argument, let us restate their critique in terms appropriate to a theory of force. (At the end of section VII, we shall consider another way of stating it.)

> *Adaptationism*$_{\mathit{force}}$ is the view that natural selection is unopposed, or only weakly opposed, by other forces of evolution, with the consequence that it always achieves the optimal result.

Gould and Lewontin think this view is mistaken because in fact natural selection is strongly opposed, indeed often thwarted, by these other forces—in particular by genetic, architectural, and developmental constraints. This invites the following question in light of item (2) in the characterization of theories of force in section II: What would happen if selection were to act *by itself*? For in order to estimate how much of an evolutionary outcome should be ascribed to nonselective constraints, and how much to natural selection, we need first to appreciate the action of the latter acting alone. In the context of the debate about adaptationism, it is clear that the supposed action of natural selection as a single factor is construed in terms of optima. The idea is that, if natural selection were to act on its own, it would achieve optima. Evolution does not always produce optima, however, because natural selection is opposed by constraints.

---

[13] Strevens points out (personal communication to Matthen) that in some cases, it may be possible to determine whether chance was involved. Strevens himself is interested in a range of cases where the i-level causes can be definitively identified. In some such cases, it might be possible to assess the exact involvement of various probabilistic causes.

[14] "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme," *Proceedings of the Royal Society of London*, CCV (1978): 581-98; reprinted in Sober, ed., *Conceptual Issues in Evolutionary Theory* (Cambridge: MIT, 1984), pp. 256-70. Page references are to the latter version.

But it is completely inappropriate to conceptualize the "power" of natural selection by specifying an optimum and asking how many obstacles it is able to overcome in achieving it, or attempting to calculate the *speed* with which natural selection would achieve an optimum acting "by itself." In population genetics, issues of speed and of overcoming constraint are addressed in the context of an underlying chromosomal structure; in evolutionary systematics, they are considered in the context of pre-existing body plans; in studies of development, they are posed as questions of heterochrony (changes of developmental sequences). In short, it is clear that natural selection acts in certain underlying causal media, and the so-called constraints are features of these media. Since natural selection cannot act without such a medium, it is not at all clear what sense can be made of the idea of natural selection acting "without the intrusion of constraints." (We shall return to this point in the next section.) If this is right, we have no way in this context of making sense of the Newtonian idea that, if an effect is to be analyzed in terms of two forces acting together, then a vector value has to be assigned to each acting independently.

The *resolution of forces* condition also fails in general in evolutionary theory. In Newton's theory, force is specifiable in absolute terms. Thus any two Newtonian forces are comparable. For example, the force produced by gravitation is comparable to that produced by electrostatic attraction: they are expressible in the same units as each other and as the force mentioned in Newton's second law of motion. It is this commensurability which allows us to estimate, for example, how an oil drop behaves when it is subject to both gravitational and electrostatic force. Coulomb's law gives us a value for electrostatic force, the law of gravitation gives us a value for gravitational force; each of these forces would act on its own if the other were absent. The law of addition of forces gives us a value for *total* force, and the latter feeds into the consequence law.

By contrast, vernacular fitness is specified in comparative terms and in restricted contexts. For example, you might learn that the optimal reproductive strategy with respect to sex determination is to produce male offspring when there are fewer males in the population, and females when there are fewer females. But this only tells you about the relative merits of strategies within a circumscribed set, with other factors held constant. The analysis does not tell you whether producing offspring of the minority sex is more or less advantageous than other fitness-relevant things you can do; there is, generally, no way of combining the effects of a good strategy in this game, with good or bad strategies in other games. For example, we have no way of calculating whether a given sex-selection strategy interacts with a given parental-care strategy, and how the fitness produced by variants of these strategies combine. This inability to add the "forces" of fitness is even more pronounced when the source laws are in unrelated domains. Suppose a certain species undertakes parental care, is resistant to malaria, and is somewhat weak but very quick. How do these fitness factors add up? We have no idea at all. The theory of probability has no general way to deal with such questions.[15]

Since we cannot add up the values produced by the source laws, the quantification of fitness in the laws of population genetics is conceptually independent of their occurrence in source laws in a way not paralleled in the Newtonian treatment of force. The overall fitness values demanded by consequence laws must be estimated statistically, that is, by looking at actual values for number of offspring, and using these actual values to estimate expected values and other statistical quantities. This independence has been remarked on by more than one philosopher of biology. "Fitness must be measured by its effects," says Alexander Rosenberg.[16] But the independence of consequence laws has often been misdiagnosed. Rosenberg thinks that it arises because there are multiple physical realizations of fitness. Sober takes a related line: "The supervenience of fitness—the fact that fitness is not a single physical property—

---

[15] Donald Davidson makes structurally the same point in "How Is the Weakness of the Will Possible?" in Joel Feinberg, ed., *Moral Concepts* (New York: Oxford, 1971), pp. 93-113. He argues (pp. 108-11) that there is no systematic way of comparing the negative value of brushing one's teeth given that one is sleepy with the positive value of doing so given that one values healthy teeth. And it follows that there is no systematic way of constructing the "all things considered" judgment on whether one should brush one's teeth.

[16] "Fitness," this JOURNAL, LXXX, 8 (August 1983) 457-73, at p. 459.

helps explain why *general* source laws are hard to come by."[17] We agree that "fitness is not a single physical property" (see section VII). But the multiplicity of realizations, whether with regard to source or consequence laws, does not mark a disanalogy between the physics and evolutionary biology: Newtonian force has multiple physical bases, too—charge, mass, collisions, and so on. The disanalogy is that, while force affords Newtonian mechanics the means to compare and add up the consequences of these diverse causes, fitness does not add up or resolve. This is why population geneticists are forced to estimate fitness by measuring population change.

V. SELECTION AND ITS SUBSTRATES

We have now argued in support of three propositions. First, there is, in general, no objective, as opposed to epistemic, apportioning of causal responsibility to selection as against drift in a concrete evolutionary history. Consequently, evolutionary theory has no zero-force law. Second, it makes no sense to think of natural selection acting "on its own," in the way that gravitation acts alone on an uncharged particle in a vacuum. So the decomposition-of-forces condition fails. Third, there is no common currency in which to compare the contributions of different evolutionary "forces." Thus, the resolution of forces does not have any counterpart in the theory. We conclude that it is inappropriate to think of evolution in terms of separate additive forces. How then shall we think of drift in relation to fitness? How shall we frame the role of constraints?

We remarked earlier that natural selection acts only in the context of certain underlying causal determinants—chromosomal structures, mode of reproduction, pre-existing body plans, developmental sequences, and so on. Instead of thinking of selection as a force acting independently of these factors, either competing with them or being reinforced by them, we shall revert to another historically familiar way of thinking about selection, namely, as a statistical trend emerging from events that occur in these "substrates." To put this somewhat fancifully, we prefer the metaphor of the hidden hand—that is, of e-level occurrences emerging from i-level events without any inherent tendency in the latter to produce them—to any conception that reifies selection as an *influence* with inherent direction. In the hierarchical-realization model that we shall now begin to construct, selection is a *formally* characterized phenomenon, a statistical property of physical substrates that possess certain metrical properties. The causally active physical properties that lie beneath this metric are different from, or even incommensurable with, the properties that form the subject matter of the theory of selection.

---

[17] *The Nature of Selection*, p. 51.

To illustrate this change of perspective, we now present two examples of the substrate dependence of selection. Recall first that Charles Darwin and Alfred Wallace hit on the principle of natural selection in an extremely general form: a heritable trait that enables its possessor to leave behind more descendants than those which lack the trait will increase its proportion in the population. Now, it turns out that, as R. A. Fisher[18] first argued in detail, the principle of natural selection is insufficient by itself to explain the facts of evolution. Fisher showed that under the system of inheritance envisaged by Darwin—"blending inheritance"—the variation in a population would die out too quickly to sustain large evolutionary changes, except under an implausibly high mutation rate, directed mutation of one sort or another, or Lamarckian inheritance (*ibid.*, chapter 1). Mendelian inheritance, in which discrete genes are handed down unchanged from parent to offspring, is required to maintain variation in the face of ecologically imposed homogenization.[19] It is not natural selection, then, that is responsible for the "origin of species," but natural selection in a Mendelian substrate. Our argument in section IV shows that this statement should not be interpreted in terms of additive forces. We shall be suggesting that a more appropriate diagnosis is in terms of certain formal properties of the Mendelian substrate.

The phenomenon of *heterosis* provides us with another illustration of substrate dependance. Sexually reproducing organisms have two genes at each chromosomal locus. Homozygotes have the same gene occur twice, heterozygotes have different genes opposite one another. Now, it sometimes happens, with respect to a particular locus, that the heterozygote is fitter (in the vernacular sense) than either homozygote, a phenomenon known, when it occurs in the plant world, as "hybrid vigor." Let us imagine an extreme case of this: the vast majority of each homozygote type aborts immediately upon conception, but most heterozygotes survive to reproduce in the normal way. If so extreme a difference of viability were to occur among *a*sexually reproducing organisms, the viable phenotype would go to fixation in short order, and the others would disappear. If it occurred with blending inheritance, the whole population would similarly settle into some nonlethal type intermediate between the two homozygotes. In the actual case, however, homozygotes continue to be produced despite the extreme negative selection against them. This is a consequence of the fact that when two heterozygotes mate, they contribute matching (particulate) genes to some proportion of their offspring, thereby producing a homozygote. Of the three substrates just considered, heterosis is manifested only in connection with sexual reproduction with particulate inheritance.

---

[18] *The Genetical Theory of Natural Selection* (New York: Dover, 1958, second edition) (the first edition was published by Oxford University Press, 1928). "No reasonable assumptions could be made by which the diminution of variance due to selection, in the total absence of mutations, would be much more than a ten-thousandth of that ascribable to blending inheritance" (p. 10).

[19] There is a significant codicil here to the so-called "tautology problem." The Darwin-Wallace proposition is a mathematical truth, consequently, some have been puzzled about how it can explain evolution. The usual response is that the antecedents of this proposition are contingent: it is an empirical fact that some heritable traits give organisms a competitive advantage. Fisher's argument shows that the Darwin-Wallace principle is insufficient to explain evolution *even under the contingent stipulation that its antecedents are satisfied.* The theory of natural selection is thus obliged to introduce considerations not mentioned in the Darwin-Wallace principle.

These analyses demonstrate the way in which selection emerges from substrate interactions summed up over ensembles. In the force model, blending inheritance would be seen as opposing selection, while particulate inheritance would be seen as reinforcing it. In addition to the infelicities noted earlier, this has the additional disadvantage that it leads us to expect only that selection would be *slower*; it gives us no reason to think that the magnitude of evolutionary change would, in the long run, be diminished. Fisher's analysis addresses the issue at a different level. It looks at *variation* within the population, and tracks the effects of this variation on the gene pool transaction by transaction, as it were. Of course, Fisher did not *literally* track individual transactions; he simply used statistical methods to sum over them. But this is equivalent to simulating ensembles of individual transactions one by one on a computer. The important point is that Fisher did not ask about the goal toward which selection is supposedly directed. Rather, he asked: 'What is the effect of random mating in a population in which some organisms have a greater probability of reproduction'? To answer this question, he followed the sequence of events through at the genic level. In his way of looking at things, selection emerges as the mathematical *aggregate* of multiple interactions at the i-level, where the latter interactions are governed by laws appropriate to the causal substrate in which these interactions occur. A similar form of analysis is used in the heterosis case. The equilibrium there is not seen as an opposition of matched forces either of which can be switched off or diminished.[20] It too arises as an aggregate of i-level occurrences.

Now, natural selection is manifested not in one, but a great variety of substrates. Biologically, it occurs in conjunction with both asexual and sexual reproduction and in the presence of varying degrees of cytoplasmic inheritance. Variations in rate of reproduction may occur at the phenotypic level, in which case effects on gene frequencies are indirect, and also at the genic level, in which case the effect is direct. Selection can be more or less determinate: the probabilities involved could approach unity, or be quite low. In possible worlds other than this one, selection may be manifested on a substrate of blending inheritance, with Lamarckian feedback, without Weismannian inviolacy of the germ line, with three or more sexes, and so on. Selection also occurs in nonbiological realms: in the economic domain, for example, as well as in "clonal selection" in the mammalian immune system, in classical conditioning,[21] and, according to some, in the propagation of theories and other cultural artifacts.[22] While the Darwin-Wallace principle continues to hold true under all these conditions, it is silent on *how* events in one generation leave their mark on subsequent generations. The properties relevant to the speed and sustainability of selection are shared by causally heterogeneous substrates, which is why these properties must be formally characterized.[23]

The multirealizability of natural selection parallels that of statistical thermodynamics. For as Lawrence Sklar[24] points out, thermodynamics has "universal applicability" across heterogeneous domains and, in marked contrast with other theories of physics, describes neither the constitution of entities nor their dynamics (*ibid.*, p. 189). Like the theory of natural selection, statistical mechanics has a formal component: this is concerned with probability distributions on large ensembles of entities interacting with each other in certain ways. This formal character accounts for its applicability to heterogeneous domains. Thus, as Sklar says, a quantity like temperature is "realized in radically different physical ways" (*ibid.*, p. 194). It is not so much a physical variable as "a way of characterizing ensembles," "a parameter appearing in some appropriate probability distribution over the microscopic states of individual systems" (*ibid.*, p. 195). Like those found in evolution, thermodynamical changes are the accumulation of a vast number of underlying events.[25] As Fisher kept emphasizing, it is statistical thermodynamics—not Newtonian dynamics—that provides the closest parallel in physics to the theory of natural selection.

---

[20] Walsh, Lewens, and Ariew argue (*op. cit.*) in a closely related argument that equilibrium states in selection, such as the one arrived at in the heterosis example, are different from those of a feather floating in air, or of an oil drop motionless under the combined effect of gravity and electrostatic force.

[21] See David Hull, Rodney Langman, and Sigrid Glenn, "A General Account of Selection: Biology, Immunology, and Behavior," *Behavioral and Brain Sciences*, xxiv (2001): 511-28.

[22] Donald E. Campbell, "Evolutionary Epistemology," in P.A. Schilpp, ed., *The Philosophy of Karl Popper*, Volume 1 (LaSalle IL: Open Court, 1972), pp. 412-63.

[23] In "Teleology and the Product Analogy," *Australasian Journal of Philosophy*, LXXV (1997): 21-37, Matthen argues that the diversity of processes in these substrates undermines the claim that there is a distinctive form of causation involved in the genesis of teleological functions, and proposes that they are based on an analogy.

[24] "The Reduction (?) of Thermodynamics to Statistical Mechanics," *Philosophical Studies*, xcv (1999): 187-202.

VI. ABSTRACT FORMULAE OF SELECTION

If selection is realized in many substrates, and has different characteristics in these substrates, what is it in general? What do the multiple realizations share? We begin now to give consideration to the formal properties of natural selection by introducing an extremely abstract characterization.

> *Li's theorem*: in a subdivided population the rate of change in the overall growth rate is proportional to the variance in growth rates.

Li's theorem can be made vivid by means of an economic example. Imagine that you put $100.00 into mutual fund *A*, and $100.00 into *B*. *A* grows steadily at 6% per annum, *B* (to your chagrin) at 2%. Now, the average growth rate of your total investment starts out at 4%. But as your subinvestment in *A* gets larger and larger relative to that in *B*, the representation of *A* in your investment account grows. This, if you like, is selection: *A* "reproduces" faster than *B*, and there are proportionately more *A*-dollars in the account. The consequence is that the average growth rate of your whole investment account becomes more and more influenced by *A* relative to *B*. For example, at the end of the first year, your investment stands at $106.00 in *A* and $102.00 in *B*; consequently, the overall growth rate is 4.04%. At the end of forty years, there is $970.00 in *A* and $216.00 in *B*, and the overall growth rate is 5.25%. Over time, the growth rate of the whole approaches the

---

[25] There are complications here. Sklar says: "Various probabilistic assumptions ...fail to have their complete grounding in either the theory of the constitution of matter or in the underlying dynamics" (*op. cit.*, p. 190). Consequently, as he says elsewhere, "Probabilistic assumptions can be built into [the] theory *at its own level*, instead of bringing [them] in at the level where atomicity and micro-mechanics are introduced"—*Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics* (New York: Cambridge, 1993), p. 23, our emphasis. We shall see at the end of section IX that somewhat analogous points can be made with regard to natural selection.

A-rate, 6%. Now, the *variance* of growth rates as between *A* and *B* was maximal at the very start of the process: as the overall mean gets closer to *A* and *A* comes to contain a greater proportion of the whole, the variance drops because the greater proportion of the account is closer to the mean. As a consequence, the increase in the overall growth rate slows down. Thus the increase in the overall growth rate tracks variance.

Anthony Edwards[26] reports that this "growth-rate theorem" was presented by C. C. Li as a "simplified version of Fisher's fundamental theorem of natural selection" (*ibid.*, p. 444). It is certainly simplified in that it does not take on board such complications as sexual reproduction, dominance, linkage, ecological change, and so on, while Fisher's theorem and its successors contain parameters that sum up the effect of such factors.[27] Consequently, Li's theorem cannot be applied to phenotypically defined subpopulations except under special circumstances: the growth rate of a trait masks the growth rate of the underlying genes, and variance in the former will not predict changes in gene frequencies. It does apply to populations of genotypes, however, though what is needed for the application is values for the growth of each genotype, or gene.[28] In effect, the theorem *defines* selection as what happens to the parts of a population when these parts have different growth rates. It implies that the faster growing subtypes increase their representation in the whole, at a speed proportionate to their variance from the mean. (Note that blending inheritance would result in a faster diminution of variance, which would damp change out faster.) This definition of selection is mathematical in nature, and independent of the particular causal laws that produce growth. (Of course, substrates with different causal laws may for that reason have different formal properties.)

---

[26] "The Fundamental Theorem of Natural Selection," *Biological Reviews of the Cambridge Philosophical Society*, LXIX (1994): 443-74, citing C. C. Li, *Population Genetics* (Chicago: University Press, 1955). Readers are referred to Edwards for a simple proof of the theorem, it is, however, quite instructive to chart the example given above in a spreadsheet program. It should be noted that the proof assumes that change in growth rates is always continuous. Patrick Maher suggests (personal communication to Matthen) that it cannot handle discontinuous changes of growth rate or discrete population growth events (for example, an annual birthing, where there are rate changes from one year to the next).

[27] See the beautiful exposition in G. R. Price, "Fisher's 'Fundamental Theorem' Made Clear," *Annals of Human Genetics*, XXXVI (1972). 129-40. Price himself has a general mathematical characterization of natural selection in "Selection and Covariance," *Nature*, CCXXVII (1970): 520-21. Steven A. Frank has a useful comparative discussion of general approaches to selection in "The Price Equation, Fisher's Fundamental Theorem, Kin Selection, and Causal Analysis," *Evolution*, LI (1997): 1712-29.

[28] Despite this predictive limitation, Li's theorem is sufficient if one is content to operate at the genic level, assuming that all relevant differences at other levels will show up there. For as Kim Sterelny and Kitcher show—"The Return of the Gene," this JOURNAL, LXXXV, 7 (July 1988): 339-61—George Williams's "book-keeping argument" implies just this. The fancier formulations of Fisher and Price add nothing to our understanding of the essential character of selection.

(Li's theorem is an abstract expression of the effects of differential growth. Taking growth rate as a surrogate for fitness, it is possible to appreciate the meaning of Fisher's "fundamental theorem": a population increases in fitness proportionate to the genetic variance in fitness among subpopulations. Note that this does not tell us much about *vernacular* fitness, that is, about the causes of growth in a subpopulation, or about progress.[29] In fact, the theorem does not predict that any given type will grow faster in time. It just tells us that the type that grows fastest will contribute most to the mean growth rate of the population.[30])

In calculating the effect of "selection pressure" on the investment account, we were dealing with a very simple case. There is no differentiation between genotype and phenotype: selection occurs directly on the kind of item being counted. Indeed, there is no "inheritance" here as such; the accumulation and continuity of "dollars" serves to preserve the effects of differential growth rates. When we deal with real biological cases, there will be effects that Li's theorem does not predict, just as there are substrate-related effects not predicted by the Darwin-Wallace formula. Li's theorem tells us nothing about causes of growth: it is a general truth about growth regardless of how it is caused. Moreover, selection is not a cause of growth (or of the change in population characteristics) in this conception; it is the mathematical aggregate of growth taking place at different rates.

VII. A HIERARCHICAL-REALIZATION SCHEME FOR SELECTION

We propose to accommodate substrate influences by conjoining various additional conditions to a basic defining formula of selection—we shall put Li's theorem in this role, though there are other candidates, such as Fisher's "fundamental theorem," Price's equation, and Richard Michod's[31] "Darwinian dynamic." The difference between the proposed approach and that of a theory of force is that a conjunction of propositional conditions does not entail, as the force analogy does, the existence of separate additive causal influences corresponding to each propositional conjunct. One can talk of consequences emerging from domains in which *p* and *q* are satisfied, without implying that these propositions correspond to separate forces.

---

[29] Fisher himself was a eugenicist for whom the vernacular associations of the concept of fitness resonated too loudly. He worried (as late as the revised postwar edition) that the fittest people—those from the upper social classes, according to him—were making themselves less fit by limiting the size of their families.

[30] It would be a mistake to think that Li's theorem (or Fisher's) depends on a growth-rate definition of *fitness*. These propositions are neutral as to whether fitness is to be defined as the "per capita rate of increase" (as Fisher understood it) or as "expected contribution of the type to the next generation" (as Sewall Wright did)—Sahotra Sarkar (personal communication to Ariew). In fact, one of us (Ariew) argues that the Fisherian definition is misleading as a definition of statistical fitness (Ariew and Lewontin, in preparation).

Accordingly, we now introduce the notion of a natural-selection *formula*:

> A natural-selection formula is one of the form (*L & C*), where *L* is the antecedent of Li's growth-rate theorem—that is, *L* posits a population subdivided by growth rates—and *C* is a *substrate specification* which states properties of population (including properties of its members or of their parts), and/or the causes of differential growth rates in these populations and their parts, and/or conditions of inheritance, development, and environmental interaction.

There is a great variety of natural-selection formulae, as the following examples demonstrate.[32] (Li's theorem is taken an implicit conjunct in each.) The Darwin-Wallace principle belongs at a very abstract level—"abstract" because the substrate specification is minimal—but less so than Li's because it is restricted to the biological domain (though Darwin was himself influenced by economic parallels). The substrate specification here is that the source of differential growth rates is variance of adaptedness, that is, of heritable traits which give organisms better ways to exploit environmental resources than those possessed by organisms with different traits, and further that these traits are inherited. Weismann proposed that changes to an organism's body could not be transmitted to its "germ plasm"; consequently, the germ plasm is passed on to subsequent generations unchanged by environmental interaction. This is an important addition to the Darwin-Wallace substrate specification, and results in a new natural-selection formula, which is distinct from Darwin's own, or Lamarck's. Again, as noted in the last section, most principles of population genetics make some assumptions about the mechanism of inheritance; the Hardy-Weinberg formula assumes particulate inheritance, sexual reproduction, and constant gene frequencies—each of these assumptions can be summarized in propositional form to yield a substrate specification in accordance with the definition given above. Fisher attempted to focus on fitness by introducing regression coefficients for fitness; assigning any definite value to these coefficients yields a natural-selection formula applicable to a particular trait in a particular organism.[33] Again, optimality analyses, which implicitly make environmental assumptions, are substrate specifications.

---

[31] *Darwinian Dynamics: Evolutionary Transitions in Fitness and Individuality* (Princeton: University Press, 1999).

[32] Hull et alia (*op. cit.*) attempt to define a formula that captures Darwinian natural selection exactly, omitting all that is extraneous and including everything essential, a formula that is neither "too broad or too narrow." We are skeptical about the value or feasibility of coming up with such a formula: our hierarchical conception undermines the idea of a unique way of specifying "gene-based evolution."

Corresponding to each of these natural-selection formulae is the set of possible histories that satisfy the formula. Let us call such a set of possible histories a *natural-selection type*. Natural-selection types are nested and overlapping. The type *SR* that corresponds to sexual reproduction will be contained in the type that satisfies the Darwin-Wallace formula; the type that corresponds to *particulate* sexual reproduction will be contained in *SR*. On the other hand, the natural-selection type that merely specifies particulate inheritance will overlap with *SR*, but not be contained in it, since there are (actual and possible) histories in which nonsexual particulate inheritance was at work, and possible (but nonactual) histories in which sexual reproduction operates through blending inheritance.

We conceive of the hierarchical relationship between more and less inclusive natural-selection formulae as similar to that between functional kinds and their realizations. Functional types such as *table*, *lock*, *camera* have concrete realizations—individual tables, locks, cameras, and so on, these being items that satisfy the defining formula of the type. Similarly, each history in a natural-selection type is a *concrete realization* of Li-selection, subject to the substrate specification *C*; a particular history will satisfy more than one such specification. (Note that an individual material object, and also a natural-selection history, could in principle belong to more than one type.) Functional types have subtypes—*refectory tables*, *combination locks*, *digital cameras*. The subtypes of natural selection are sets of histories that satisfy a particular substrate specification. These are *kinds* of Li-selection—Li-selection with Mendelian inheritance, with sexual reproduction, and so on. These types have subtypes, too. This is why we call our model a *hierarchical-realization scheme*. Note that the novelty of our proposal does *not* lie in our making selection multirealizable; earlier, we quoted Rosenberg and Sober as already committed to this. The new feature in the scheme outlined above is its hierarchical arrangement, with various factors taken to be explanatory of evolution figuring in the substrate specifications.

---

[33] See Price, "Fisher's 'Fundamental Theorem'," p. 130. Fisher counted dominance relations and other chromosomal effects as "environmental." Sterelny and Kitcher make the same move (*op. cit.*).

We arrive at an adequate explanation of the evolution of a biological phenomenon by subsuming it under the *most specific* formula that applies to it, that is, the formula that posits *all* the substrate factors relevant to it. The probability of the target phenomenon is estimated relative to the histories that constitute the corresponding *least inclusive* natural-selection type. We understand why the phenomenon came about by comparing this probability with those yielded by natural-selection formulae which impose relevantly different substrate specifications. For instance, we understand why a deleterious hereditary condition like sickle-cell anaemia was not eliminated, by comparing the probabilities in its natural-selection formula with those with relevantly different ones—ones in which malaria was not a factor, ones in which reproduction is not sexual, and so on.

This paves the way to a new interpretation of adaptationism. It consists not so much in overestimating the "power" of natural selection to overcome the other determinants of evolution, but in overestimating the explanatory powers of natural-selection formulae that lean too heavily on optimality analyses concerning traits, omitting specification of other factors. We noted before, for instance, that evolution would be minimal if there were blending inheritance rather than particulate. For this reason, the Darwin-Wallace formula fails to predict the consequences of differences in vernacular fitness—it lacks an adequate substrate specification. Darwin overestimated the explanatory power of natural-selection formulae that omit specific mention of inheritance and mention only optimality analyses. To this extent, he was an adaptationist.

VIII. THE DISAPPEARANCE OF VERNACULAR FITNESS

Consider now how to approach questions of interaction among evolutionary factors. How, for instance, does sexual reproduction interact with selection? Where Sober's force model posits an additional force for sexual reproduction, we capture its effects by moving from a natural-selection formula in which sexual reproduction is not specified to one in which it is. Separability is not implied: a formula that fails to specify a mechanism of inheritance does not, for that reason, specify selection *without* inheritance. When we consider evolutionary effects that do not depend on sexual reproduction, or when we wish to consider an effect in abstraction from mode of reproduction, we might ascend to a more inclusive type that includes, for instance, *all* forms of inheritance (possible and actual). In this more inclusive selection type, the growth rates are averaged out over all forms of inheritance. There is no selection in this or any other type in the absence of some sort of continuity that could be considered a fill-in for inheritance.

In this way of looking at things, the distinction between evolution (the total change of gene frequencies due to all causes), and natural selection (the portion of evolution due to differences in competitive advantage) is unmotivated. Natural selection is, as Li's theorem tells us, the aggregative result over time of differential growth rates in a population. These growth rates are explained by considering *all* of the factors posited by the most specific relevant natural-selection formula, competitive advantage acting in concert with all the others. In histories that conform to this formula, certain trends get established at the e-level as accumulations of multiple concrete events—births, deaths, mate choices, as well as events at the cellular and molecular level. These trends constitute natural selection. There is no difference between these trends and evolution. Further, since these trends constitute total change from generation to generation, predictive fitness is their measure, not the supposed partial value that some posit as vernacular fitness. Like the distinction between natural selection and evolution, the distinction between vernacular fitness (an organism-type's overall competitive advantage traceable to heritable traits), and predictive fitness (its expected growth rate, or the propensity underlying it), is unmotivated.

Some might resist this conclusion on the grounds that actuarial practice regularly apportions causal responsibility to various factors. (After all, this is what multivariate analysis is all about.) In responding to this concern, we want clearly to acknowledge that it *is* legitimate to ask, *in a statistical sense*, how much of the causation of B is due to competitive advantage, though as we noted in sections III and IV, one needs to be prepared for a degree of indeterminacy in the answer, since the composition and decomposition of probabilities is mathematically undefined in many cases. This question is similar to the questions asked when determining insurance premiums. For example: How much does being a male youth contribute to road accidents in which male youths are involved? There is some sort of answer to this question in statistical correlations. But this does not imply that being a male youth operates in an ontologically separate way in these accidents. This is clearest when we think about the individual events. An eighteen-year-old male is involved in a collision at night, in the rain, after drinking three pints of beer. Are these factors, specified in this way, separable influences on the outcome? We think not: the same goes here as for the factors involved in cases like (C1) and (C2) in section III.

IX. PROCESS VERSUS TREND IN EVOLUTION

We turn now to some considerations of causality and process. We have been insisting that natural selection is a mathematical aggregate of individual events. This seems to imply that it is not a causally connected process. The increase of fitness in a population by means of natural selection is, as we see it, a temporal sequence of aggregates, a running tally of lower-level events; it is not a process within which the earlier events cause the later. The same holds for the spread of adaptive features through a population. This may seem counterintuitive. It helps to consider parallels from thermodynamics. Is increase of entropy a process? What about the transfer of heat from high to low temperature?

Let us start by considering the case of a pot of water brought to a boil on a stove. At a gross phenomenal level, this seems like an orderly process. Heat flows from stove to liquid, the liquid circulates, carrying heat with it; gradually, it comes to a boil. Closer examination shows, however, that the transformation is not an orderly one at all. This is most evident perhaps in the actual phase transition, which is highly jerky. Parts of the liquid heave up and down; bubbles form more or less randomly. When the liquid is actually boiling, its surface is chaotic. A microscopic examination would display similar disorder at earlier stages. The energy transfer from the bottom of the pan to the top is disorderly, with myriad local exceptions (for example, adjacent small regions with unequal temperature with no mutual heat transfer), reversals (energy flow from low- to high-temperature regions), and other fits and starts. Such *discontinuities* violate the spatiotemporal continuity required of *fundamental physical processes*, which are strictly law governed. To halt, reverse, delay, or accelerate a fundamental process takes energy and work. But in the pot of water such changes of direction occur spontaneously. This shows that heat flow is not a fundamental process. The same point can be made (and historically was made) by reference to *reversibility*.[34] Played backward, a fundamental physical process conforms to the laws of physics; heat transfer does not. You cannot get heat to flow spontaneously back from the boiling water into the stove top by reversing thermodynamic determinants, as you can get a particle to retrace its path by reversing all the momenta in its ensemble.

---

[34] An accessible account of reversibility can be found in P. W. A. Atkins, *The Second Law* (New York: Freeman, 1984).

In physics, discontinuous processes of the above kind are treated as aggregates of fundamental processes. Statistical mechanics handles conductive and convective heat transfer stochastically. The pot of water consists of a large number of molecules in random motion. Let us say that a *mechanical* state description of such a system is the set of position and momentum vectors for each particle it contains. Each *thermodynamic* state description of the system—the set of values for its energy, temperature, and other such thermodynamic variables—supervenes on a mechanical state description thereof. But this supervenience relation is *one-many*: that is, the same thermodynamic mechanical state of a given system supervenes on many different mechanical states. Now, some thermodynamical states comprehend a larger number of distinct underlying mechanical states than others: those which comprehend more are "more probable," those which comprehend fewer are "less probable." The basic principle of statistical thermodynamics is that less probable thermodynamic states give way in time to more probable ones, simply by the underlying entities participating in fundamental processes. This explains why statistical thermodynamics has "universal applicability," as Sklar says (see the end of section V, above): it is silent about the constitution of entities and their dynamics, and mathematically demonstrates a phenomenon that occurs regardless of these physical details. The heat flow described above is stochastic in nature; it is a trend predicted by the mathematics of large numbers, not a fundamental process governed by the specificities of physical law.

The same is true of evolution. The discontinuities of natural selection are much more evident to the observer than those of thermodynamics, gross enough, in fact, to be recordable by a careful observer.[35] These discontinuities show that natural selection, like thermodynamic change of state, is a time-asymmetric statistical trend instantiated by populations. In an important sense, the mathematical-statistical treatment of natural selection eliminates evolutionary process. That is, it shows that earlier and later states of population fitness or adaptedness in an evolutionary process should not be considered stages in the unfolding of a fundamental process.[36] This is what the hierarchical-realization scheme is intended to capture: natural selection is the consequence of heterogeneous processes in substrates sharing only a formal structure. Let us define a *stochastic property* as one that belongs to ensembles as a *mathematical* (note: not *nomic*) consequence of the i-level properties of individuals in that ensemble. Further, define a *trend* as a change of an ensemble over a period of time with respect to one or more of its stochastic properties. The claim that we want to make is that while predictive fitness values are predictors of trends in populations, and may thus be considered *probabilistic causes*, they are not causes in the sense appropriate to fundamental processes.

---

[35] For a famous and vivid study of discontinuities and reversals in selection, see Peter Grant, *Ecology and Evolution of Darwin's Finches* (Princeton: University Press, 1986).

[36] In "Chasing Shadows: Natural Selection and Adaptation," *Studies in the History and Philosophy of Biology and the Biomedical Sciences*, XXXI (2000), 135-53, Denis Walsh

We should distinguish this conclusion from two other propositions with which it might be confused. First, some will find that our confining fundamental process to the i-level, and insisting that e-level transitions are trends, reminds them of the idea that in the theory of natural selection, probabilities are *epistemic* only, used because we are ignorant of the individual deaths and births that constitute the i-level reality of natural selection. In fact, we are not sympathetic to the latter notion; we do not regard statistical theories merely as devices to deal with ignorance. The statistics employed in thermodynamics *explains* something about the nature of heat and work, something we would not understand by having a history of molecular interactions in a pot of water, even assuming that our minds could comprehend that kind of detail. Similarly, the statistics of natural selection tells us something deep about the patterns instantiated in diverse biological histories. By appreciating these patterns, we come to understand something that we are not able to see when we are given the full biographical details of organisms in diverse populations.[37] The theory of evolution is a historical science in the sense that it tries to retrodict significant events in the individual paths traced by actual species. Population genetics, however, is concerned with the formal characterization and mathematics of all such histories, taken as a group. The

---

argues, on grounds independent of those given here, that heat transfer and natural selection are "pseudo-processes."

[37] This is so even when the mathematical reason for trends is not explicitly stated. Consider the computer simulations used by Peter Danielson in *Artificial Morality: Virtuous Robots for Virtual Games* (New York: Routledge, 1992), and Brian Skyrms in *Evolution of the Social Contract* (New York: Cambridge, 1996). Ariew discusses the explanatory value of statistical patterns in "Are Probabilities Necessary for Evolutionary Explanations?" *Biology and Philosophy*, XIII (1998): 245-53.

interest of these patterns is not simply historical reconstruction, any more than it is in thermodynamics.[38]

Second, some may think that we are asserting that, if a class of properties *S* supervenes on base properties *B*, then since all changes in properties *S* are wholly determined by properties *B*, there are no genuine causal relations at level *S*. In fact, we have not relied on the supervenience relation between e-level properties and i-level properties in making our point. We have distinguished two kinds of causal relations, fundamental and stochastic. We concede that stochastic causation occurs at the *S* level,[39] but deny that process causation occurs at this level. This denial is made *on grounds of discontinuity and irreversibility*, not of supervenience.

The view that natural selection is a cause of evolution is becoming increasingly fashionable. Some place the recently heightened interest in developmental constraints and morphogenetic process in this context; they say that these influences on organismic traits should be given more prominence relative to natural selection than they have been given hitherto. In our opinion, this is misguided. It relies on the idea that fitness and natural selection are separable causes of evolution. We have argued for a hierarchical-realization model of selection, in which fitness is a probability defined on the class of selection histories that meet all the relevant explanatory conditions. Further, we have argued that the nature of causal relations is different at the e-level, where natural selection occurs, and the underlying i-level. Fitness and natural selection have no reality except as accumulations of more fundamental events. In our conception, it makes dubious sense to hold evolution is different from natural selection, much less that it is caused by it.

Having said this, we do want to acknowledge that the construction of a genuinely all-inclusive conception of natural selection in the biological domain is extremely challenging. Sklar's discussion of thermodynamics is once again revealing here. Speaking of temperature, he says:

> Systems of wildly divergent physical kinds can all have equilibrium states and, when allowed to transfer energy to one another, be in equilibrium with respect to each other. The moving molecules in a blob of matter, for example can form a system in equilibrium with electromagnetic radiation. So we are required to think of these systems as sharing common temperatures.[40]

---

[38] Ariew makes similar points in "The Ultimate/Proximate Distinction" (in preparation).

[39] Here we concur with Sober, "Physicalism from a Probabilistic Point of View," *Philosophical Studies*, xcv (1999). 135-74, especially pp. 145-48.

[40] "The Reduction(?) of Thermodynamics to Statistical Mechanics," p. 194; see also footnote 25 above.

Something like the same problem arises in the theory of natural selection because of the variety of levels at which selection events occur—these range all the way from that of molecules to that of social groups. An example of this lies in the importance of both genetic selection and cultural selection in human evolution.[41] Here, as in Sklar's example, selection is occurring simultaneously in two quite different substrates. The question is: How do we understand the interaction of these two substrates? What is the nature of the equilibria that is reached in these cases? The imposition of such crossing relations on our hierarchical-realization scheme is no small challenge. It may well be that in the last analysis, fitness involves sui generis probabilities estimated not by causal, or even mathematical, analysis, but by census.[42]

MOHAN MATTHEN

University of British Columbia

ANDRÉ ARIEW

University of Rhode Island

---

[41] See P. E. Griffiths and R. D. Gray, "Developmental Systems and Evolutionary Explanation," this JOURNAL, XCI, 6 (June 1994): 277-304; and Susan Oyama, *The Ontogeny of Information. Developmental Systems and Evolution* (New York: Cambridge, 1985).

[42] The differences between physical and statistical ways of thinking about factors relevant to evolution have been investigated independently by Denis Walsh and Matthen. Ariew, who has been working with both, brought about considerable convergence both by cross-communicating and by his own critical contributions. The present article and Walsh, Lewens, and Ariew (*op cit.*) are products of this process. But these articles are almost completely independent with respect to argumentation and even, to a surprising extent, motivation.

CAUSATION: INFLUENCE VERSUS SUFFICIENCY*

David Lewis[1] has produced a new theory of singular-event causation. He still maintains that C is a cause of E if and only if C and E are distinct actual events linked by a chain of causal dependence. But his account of direct causal dependence has been revised. According to the *old theory*,[2] E is causally dependent on C if and only if E would not have occurred if C had not occurred. According to the *new theory*, E is causally dependent on C if and only if C *influences* E. C influences E if and only if

> ...there is a substantial range C_1, C_2...of different not-too-distant alterations of C (including the actual alteration of C) and there is a range E_1, E_2...of alterations of E, at least some of which differ, such that if C_1 had occurred E_1 would have occurred, and if C_2 had occurred E_2 would have occurred, and so on.[3]

An *alteration* of an event E is officially defined as "either a very fragile version of E or else a very fragile alternative event that is similar to E, but numerically different from E" (*ibid.*, p. 188). It transpires, however, that the *absence* of the actual alteration of E also counts as an alteration of E.

The new theory goes some way toward solving the problems that confronted the old theory, but, I shall argue, not far enough. Better results can be achieved on an alternative approach, according to which a cause is a necessary part of a sufficient condition. Conditionhood, however, should be understood in terms of counterfactuals, not the instantiating of the laws of nature. Lewis's fundamental thesis was correct: causation can be analyzed in terms of counterfactuals.

I. LATE CUTTING

The first supposed advantage of the new theory is that it solves the problem of *late cutting*. Suzy and Billy throw rocks at a bottle. Suzy's rock reaches the bottle a split second before Billy's, and it shatters. If Suzy had not thrown, the bottle still would have shattered. So, under the old theory:

(1) Suzy's throw caused the shattering.

---

* I am indebted to Wylie Breckenridge, Christopher Hitchcock, David Lewis, and Jonathan Schaffer for helpful comments and suggestions.

[1] "Causation as Influence," this JOURNAL, XCVII, 4 (April 2000): 182-97.

[2] "Causation," this JOURNAL, LXX, 17 (October 11, 1973): 556-67.

[3] "Causation as Influence," p. 190.

0022-362X/02/9902/84–101

seems to be false, contrary to common sense. For the shattering does not seem to be counterfactually dependent on Suzy's throw, or on any event in the process which (intuitively) links her throw to the shattering.

The new theory says (1) is true. C_1 is the actual alteration of Suzy's throw, C_2 is the null alteration (she does not throw at all), E_1 is the actual alteration of the shattering, E_2 is the slightly later shattering that Billy's throw would have produced. So Suzy's throw influenced the shattering. (We could embroider the case, adding details about small variations in the time or manner of Suzy's throw, or chains of intermediate causes, but the new theory gives the right answer as long as we have the facts stated, which are of a pattern common to all cases of late cutting.)

In 1986, Lewis[4] considered and rejected a somewhat similar response to the late-cutting problem, the *extreme-fragility response*. This stipulates that events have extremely rich essences: no event could have occurred at any other time or in any other manner than it actually did. Under the old theory with extreme fragility, (1) is true because

(2) If Suzy's throw had not occurred, the shattering would not have occurred.

is taken to be true: although the bottle would have shattered, it would have been a different shattering.

Lewis says the new theory is better than the extreme-fragility response for the following reason. In our explicit talk about event essences, we are sometimes undecided. (Jonathan Bennett's[5] example (*ibid.*, p. 55): 'My handwave would not have occurred if I had waved with the left hand instead of the right.' True or false?) The new theory does not rely on any assumptions about event essences.

In 1986, Lewis had noted a related (but bigger) objection to extreme fragility. It flatly conflicts with a lot of our explicit talk about event essences. We say (and we are not a bit undecided about it): "The shattering would still have occurred, only a bit later, if Suzy had not thrown." But, Lewis said, "this is not a decisive objection. The standards that apply within the analysis of causation might differ from those that apply in explicit talk."[6] I think this was a sensible reply to the objection. It answers the new objection to extreme fragility as well. The new theory is no great improvement on this score.

---

[4] "Postscript E to 'Causation'," in *Philosophical Papers*, Volume II (New York: Oxford, 1986), pp. 193-212.

[5] *Events and Their Names* (Indianapolis: Hackett, 1998).

[6] "Postscript E to 'Causation'," *Philosophical Papers*, Volume II, p. 198.

The reason Lewis rejected extreme fragility in 1986 was the *problem of spurious causation:* it makes many causal sentences that are intuitively false come out true. One example concerned Boddie: he eats a big dinner, and as a result the poisoned chocolates take longer to kill him. Extreme fragility makes

(3) Boddie's eating a big dinner caused his death.

true.

This is a problem for the new theory, too. Boddie's dinner influenced his death, by delaying it.

Lewis's response is that influence, and hence causation, is a matter of degree. In a context where C_1 has a substantial influence on E and C_2 has a negligible influence (by comparison), we neglect C_2's influence and count 'C_2 caused E' as false. In an ordinary context, where the poisoned chocolates are salient, Boddie's dinner had a negligible influence on his death. (Without the dinner, he would have died *minutes* earlier; without the chocolates, he would have died *years* later.) So we count (3) as false.

This is not an adequate response. Even among Lewis's own examples of spurious causes, there are some which have an influence that is substantial in any context. The doctor's treatment delays the patient's death by many years. Therefore, according to the new theory:

(4) The doctor's treatment caused the patient's death

is true. It is true according to extreme fragility as well. "That will not do," Lewis said, in 1986.[7] It still will not do.

(Of course, a treatment that saves a patient's life *can* also cause his death—for example, it might introduce some slow-acting infection. What is unacceptable in Lewis's position is that, simply by saving the patient's life, the doctor *automatically* counts as causing his death.)

To put late cutting to one side, I shall until further notice consider only examples where we can agree (under any reasonable standards of event essence) that the purported effect would not have occurred at all, if it had not happened exactly when and how it did.

II. THE ESSENCE OF THE CAUSE

The second supposed advantage of the new theory is that it solves what we may provisionally call the *problem of the essence of the cause.* This problem for the old theory was not, I think, discussed by Lewis before 2000.

---

[7] "Events," in *Philosophical Papers*, Volume II, pp. 241-69, here p. 250.

He begins by pointing out that the meaning of 'If *C* had not occurred, *E* would not have occurred' depends on what would count as *C* not occurring—a matter about which we are, in some cases, indecisive. He continues:

> That makes a problem. What is the closest way to actuality for *C* not to occur? It is for *C* to be replaced by a very similar event, one that is almost but not quite *C*, one that is barely over the border between versions of *C* itself and its nearest alternatives. But if *C* is taken to be fairly fragile, then, if almost-*C* occurred instead of *C*, very likely the effects of almost-*C* would be almost the same as the effects of *C*. So our causal counterfactual will not mean what we thought it meant, and it may well not have the truth value we thought it had. When asked to suppose counterfactually that *C* does not occur, we do not really look for the very closest possible world where *C*'s conditions of occurrence are not quite satisfied. Rather, we imagine that *C* is completely and cleanly excised from history, leaving behind no fragment or approximation of itself. One repair would be to rewrite our counterfactual analysis, or add a gloss on its interpretation, in order to make this explicit.[8]

This is pretty puzzling. What Lewis *says* is that the problem is a problem for his analysis of *counterfactuals*, not for the analysis of causation in terms of counterfactuals. The problem is that when asked to evaluate 'If *C* had not occurred,...', we do not do what Lewis's analysis of counterfactuals says we do: we do not look at the closest possible world in which '*C* does not occur' is true. If we did, Lewis says, we would often judge 'If *C* had not occurred, *E* would not have occurred' to be false when *C* really causes *E* (because the closest world in which *C* does not occur has almost-*C* instead); but we do not.

I presume this is just a slip. Lewis *means* that '*C* caused *E*' is often true when 'If *C* had not occurred, *E* would not have occurred', understood literally (and, hence, according to his theory of counterfactuals), is false, because 'If *C* had not occurred, *E* would have occurred' is true instead. So he must change the analysis of 'cause'.

But *why* does Lewis think that, when *C* causes *E*, 'If *C* had not occurred, *E* would have occured' is often true—either intuitively, or according to his theory of counterfactuals—and even supposing *C* to be fairly fragile? His analysis of counterfactuals does *not* say that the closest world in which *C* does not occur is usually one in which *C* is replaced by a very similar event. Closeness of worlds is a matter, rather, of smallness and lateness of miracles. (Remember the line: Lewis does not want his theory to imply that, if the line had been longer, it would have been less than 0.0001" longer.) Intuitively, it seems that when we consider what would have happened if *C* had not occurred, we usually consider a variety of possibilities, and it is *not* true that something very similar to *C* occurs in all of them. In Lewis's terms, the closest not-*C* worlds usually seem to be a mixed lot, with *E* occurring in some and not in others. I agree that we shall often judge that 'If *C* had not occurred, *E* would not have occurred' is false. But it seems implausibly strong to claim that we shall often judge that 'If *C* had not occurred, *E* would have occurred' is true. And unnecessarily strong: all that is needed for there to be a problem for the old theory is that 'If *C* had not occurred, *E* would not have occurred' is false.

---

[8] "Causation as Influence," p. 190.

The problem Lewis seems to be groping for, then, is this. Often when '*C* caused *E*' is true, 'If *C* had not occurred *E* would not have occurred' is false because *some* alternatives to *C* would still have produced *E* (some closest not-*C* worlds are *E*-worlds). It is a problem I raised in 1995.[9] My main example was "Two Buttons": I pressed button *A* and Victim got a shock; if I had pressed *B* he still would have; if I had pressed neither he would not. 'My pressing *A* caused the shock' is true, but 'If my pressing of *A* had not occurred the shock would not have occurred' is false. All that is true is that, if I had not pressed *A*, the shock *might* not have occurred.

The new theory solves this problem. My pressing *A* influences the shock because there is an alteration of the pressing, the null alteration, such that, if it had occurred, the shock would not have occurred.

The solution I proposed in 1995 was simpler: replace the 'would' in Lewis's causal counterfactual by 'might'. The old theory said: *E* is causally dependent on *C* if and only if *E* would not have occurred if *C* had not occurred. I suggested: ...if and only if *E* *might* not have occurred if *C* had not occurred. It is the difference between

(a) If *C* had not occurred, *E* would not have occurred.

and

(b) Not-(if *C* had not occurred, *E* would have occurred).

What makes my pressing *A* a cause of the shock is simply that it is *not* the case that the shock would have occurred whatever I had done.

---

[9] "Lewis on Causal Dependence," *Australasian Journal of Philosophy*, LXXIII (1995): 129-39.

III. TRUMPING

The third supposed advantage of the new theory is that it solves the *problem of trumping* (so-called).[10] The laws of magic say that what happens at midnight corresponds to the first spell cast in the preceding twenty-four hours. At noon, Merlin casts a spell to turn the prince into a frog. At 1:00 pm, Morgana casts a spell to turn the prince into a frog.

(5) Merlin's spell caused the prince's frogification.

is clearly true, but false according to the old theory.

The new theory solves the problem. Merlin's spell influences the frogification, since if Merlin had instead cast a spell to turn the prince into a goat, say, the frogification would not have occurred.

My 1995 'might' analysis solves the problem as well. Merlin's spell caused the prince's frogification because it is not the case that the frogification would have occurred whatever Merlin had done: if Merlin's spell had not occurred the frogification *might* have not occurred, because Merlin might have cast the goat spell instead.

The 'might' analysis was not supposed to solve *all* problems for the old theory. It was just intended to illustrate one small step toward a better theory. It was just good luck that it anticipated the trumping problem.

I have been taking the problem to be, how to get (5) to turn out true. Lewis thinks the old theory's verdict on

(6) Morgana's spell caused the prince's frogification.

was right—that is, it is false. The new theory also says it is false. So does the 'might' analysis. My own intuition is that (6) is *true*. It is an intuition of "intrinsicness": the (one-step) process from Morgana's spell to the prince's frogification runs to completion exactly as it would have done in the absence of any Merlin spell, and it would certainly have been a causal process in that case. Elsewhere, I have given arguments against the idea that causation is intrinsic, but they had to do with "double prevention": in the present kind of case, the intuition seems to be undefeated. The so-called trumping case (with no intermediate events) is one of overdetermination, not preemption.

Jonathan Schaffer (*op. cit.*, p. 165) argued that (5) was true and (6) false simply because the laws of magic say that the *first* spell of the day comes true. But this was a bad reason. If it is a law that

---

[10] Jonathan Schaffer, "Trumping Preemption," this JOURNAL, XCVII, 4 (April 2000), 165-81.

(i) The first spell always comes true.

then it is also a law that

(ii) When nonequivalent spells are cast the first comes true, but when only equivalent spells are cast the last comes true.

In fact, it is the same law. If p is logically equivalent to q, it is a law that p if and only if it is a law that q. But if we reasoned in Schaffer's way with this alternative formulation of the law, we would conclude that Morgana's spell caused the frogification and Merlin's did not. The moral is that the laws inform our causal judgments, not in Schaffer's very direct way, but by telling us what would have happened if things had gone differently.

(Schaffer does make some attempt to answer this objection. He argues, assuming the supervenience of laws on actual facts à la Ramsey-Mill-Lewis (RML), that if (i) is a law then (ii) is *not* also a law. According to RML, "L is a law of nature if and only if L is a theorem of the axiomatization of the facts that best balances simplicity and strength" (*op. cit.*, p. 171). Schaffer points out that (ii) is less simple than (i). But this is beside the point. Simplicity is a criterion for the *axioms* of our system of the world, according to RML; but the *theorems* can be of unlimited complexity. Clearly, (ii) is a *theorem*—a law—if (i) is.)[11]

Right or wrong, the new theory and the 'might' analysis give the *same* verdicts, on the cases discussed so far. One might suspect that the new theory is *equivalent* to the 'might' analysis, under our current sidelining of late cutting. Is 'For some alteration C_1 of C, if C_1 had occurred E would not have occurred' not equivalent to 'If C had not occurred, E might not have occurred'? No.

(i) What might have occurred if C had not occurred depends on what else could have happened *at that point*. It is *not* true, for example, that if Morgana had not cast her frog spell she might have cast a goat spell *earlier*. For, since Morgana *did not* cast any earlier spell, it is true (in the standard, nonbacktracking sense of counterfactuals, which is the sense relevant to causation) that, if Morgana had not cast the frog spell at 1:00 pm, she *would* not have cast any earlier spell.

---

[11] One might, I suppose, assume that a law statement can itself incorporate causal terms. Then 'It is a law that the first spell always *causes* what happens at midnight' would perhaps not imply 'It is a law that when only equivalent spells are cast the last *causes* what happens at midnight'. But such an assumption would be both implausible and question begging. There is at one point a suggestion that it is indeed Schaffer's assumption: "Clearly, Merlin's spell (the first that day) is the cause of the prince's becoming a frog and Morgana's is not, because the laws say that the first spells are the consequential [*sic*: he must mean 'efficacious'] ones" (*op. cit.*, p. 165). But everywhere else Schaffer takes it that a law statement is of the form 'It is a law that p', where p itself contains no causal or modal terms.

Lewis allows, however, that an "alteration" of *C* may have a different time of occurrence. He needs to allow that an alteration of the *effect* may have a different time of occurrence, to solve the problem of late cutting. To treat the cause differently would be "one-sided" (and, Lewis evidently thinks, therefore bad).

But allowing for earlier alterations of the cause creates problems. Consider a variant of the trumping case in which Morgana's spell is a goat spell (at 1:00 pm). There is no doubt whatsoever that in *this* case 'Morgana's spell was a cause of the prince's frogification' is false. But one alteration of Morgana's spell is a goat spell at 11:00 am. And if Morgana had cast this spell the prince's frogification would not have occurred. So why did not Morgana's 1:00 pm goat spell influence, and hence cause, the frogification?

The "not-too-distant" in Lewis's definition of influence seems intended to exclude such cases: a possible 11:00 am spell is too distant from the actual 1:00 pm spell to be relevant to the actual spell's influence.

If Merlin's spell had preceded Morgana's by less than a minute, a possible Morgana spell a minute earlier than her actual spell would be too distant. "Too distant" is obviously intended to be context relative. But relative to *what* in the context? Consider a case in which the laws of magic state that a goat spell takes precedence over a frog spell as long as it is cast not more than fifty-nine minutes later than the frog spell. (Merlin's and Morgana's actual spells are still a 12:00 pm frog spell and a 1:00 pm goat spell.) Now is a possible Morgana goat spell at 12:58 pm too distant from her actual spell? If not, 'Morgana's 1:00 pm goat spell was a cause of the prince's frogification' is true; this is still clearly anti-intuitive (although 'Her not casting a goat spell at 12:58 pm was a cause' may be true). If two minutes earlier *is* too distant, then "too distant" means...what? It begins to look as if it just means distant enough to give the wrong answer.

(Lewis also stipulated that an alteration of *C* must be "similar" to *C*. But as far as similarity in respects other than time of occurrence is concerned, he seems simply to ignore that stipulation, and the "not-too-distant" in the definition of influence. For example, Merlin's goat spell might be an intrinsically *very* different performance from his frog spell. And surely, the null alteration, the absence of any spell-casting on Merlin's part, would not be at all similar to his actual spell.)

(ii) C may influence E even when E definitely would have occurred in the absence of C, on any ordinary standards of event essence. The influence may come from the fact that E would not have occurred if C had occurred in a different manner. For example, a concert once caused a riot in virtue of its extreme brevity. I accept 'The concert caused the riot'. Intuitively, my reason is that the riot would not have occurred if the concert had been longer—if it had not occurred at all, the customers would have reacted just as they did in actuality. But unless we accept extreme fragility, and say that the concert would not have occurred if it had been longer, 'The concert caused the riot' is false on the 'might' analysis. Lewis's new theory gets this one right. (Actually, I think that accepting extreme fragility *in relation to the cause only* is a good idea: see below.)

So much for the supposed advantages of the new theory. But there are several other long-standing problems for the old theory which the new theory does nothing to solve.

IV. OVERDETERMINATION

Neuron E is stimulated by the simultaneous firing of C_1 and C_2, and fires exactly as it would have if C_1 or C_2 had fired alone. The new theory, like the old, denies that the firing of C_1 was a cause of the firing of E, contrary to common sense. (I assume here that Lewis would have to say that, since the firings of C_1 and C_2 were simultaneous, even a slightly earlier possible firing of C_1 would be too distant from the actual firing, so that there is no when-on-when dependence.)

Lewis's response was to reject the deliverances of common sense (as I take them to be): the case should be regarded as "spoils to the victor." Well, I disagree; but you will have to consult your own pretheoretic intuitions.

V. EARLY CUTTING WITH PREEMPTING DIRECT CAUSE

We assume that only one spell works for Merlin, the frog spell: if he attempts to cast any other spell, nothing happens. He casts this spell. Morgana does not cast any spell, but would have cast a frog spell if Merlin had not done so. Clearly, Merlin's spell was the cause of the prince's frogification. But Lewis's new theory, like the old, says it was not.

(I call this *early cutting* because (i) the process from the preempted cause (whatever it was which would have been a joint cause, with Merlin's inaction, of Morgana's spell—call it *Morgana's preparedness*) to the effect, via a Morgana spell, is cut short; and (ii) it is cut short, not by the effect, but by an event in the (one-step) preempting process—that is, by Merlin's spell. Lewis characterizes early cutting as "easy cases in which we have, if not direct counterfactual dependence of the effect upon the preempting cause, at least *stepwise dependence*."[12] But I think this is just because he had not considered cases like the one above.)

VI. TRANSITIVITY

Lewis's new theory, like the old, implies that causation is transitive. But there are cases where many competent speakers are prepared to say that *C* caused *D*, *D* caused *E*, but *C* did not cause *E*.

Lewis's response seems to be that such speakers are confused; and he makes some suggestions as to the causes of the confusion. I do not think this strategy could possibly work. According to Lewis, what the speakers in question *mean* is that there is a chain of causal dependence going from *C* to *D*, a chain from *D* to *E*, but no chain from *C* to *E*. But hardly anyone could be so confused as to think *that*!

Some may think my comments reflect an excessively naive view of conceptual analysis. We cannot really be expected to *know* what we mean, they say. So let us pursue the question further, whether '*C* caused *D* and *D* caused *E*' entails '*C* caused *E*'.

Consider a standard apparent counterexample to transitivity, "The Switch": the engineer throws the switch so that the train goes down the right-hand track instead of the left-hand track; but the tracks converge, and the train still arrives at its destination. Initially, it seems clear that the switching was a cause of the train's passing over the mid-point of the right-hand track (say), that its passing over this point was a cause of its arrival at its destination, but that the switching was not a cause of its arrival.

But this intuitive response can be reversed. The most effective kind of argument, in my experience, goes as follows. Consider the *unused* track, the left-hand track, and consider some appropriate event *X* involving it—its existence/construction/repair/.... If neither *X* nor the switching had occurred, the train would not have arrived. So, between the two of them, *X* and the switching caused the arrival. Now *which one* caused the arrival—*X* or the switching (or both together)? At this point, nearly everyone retracts his original intuition and says, "Well, it must have been the switching that did it—*X* clearly contributed nothing."

I have used this kind of argument before. Before considering its significance for transitivity, let me say something about its power to change our intuitions. John Collins[13] has analyzed the matter in some detail; but I think he has got the main point slightly wrong.

---

[12] "Causation as Influence," p. 184.

The focal case is "The Catcher": I catch a ball heading toward a brick wall and, beyond that, a window. Did the catch cause the ball to not hit the window? Our initial reaction is "No," but this is reversed when it is pointed out that the catch and the wall between them caused the ball not to hit the window, and the wall contributed nothing. (I mean, of course, some appropriate *event* involving the wall.)

Collins usefully contrasts this case with one in which the wall is replaced by a second catcher, who would have caught the ball if I had not. Here the argument for "Yes" springs to mind quite naturally, and most speakers see right off that my catch did cause the ball not to hit the window. Why the difference?

Collins says that the supposition of the absence of the wall is "more far-fetched" than the supposition of the second catcher's failure to catch the ball. What does he mean by "far-fetched"? Does he mean that the proposition 'The wall was not there' is less *credible* than the proposition 'The second catcher would not have caught the ball if it had come to him'? Obviously not: both propositions are known to be false. Perhaps it is the *supposing* of the proposition, rather than the proposition itself, that is supposed to be "far-fetched": Collins suggests that it is more of "a stretch"—more difficult?—to suppose away a wall than an ability to catch. But this will not work either: in a case where there is no catcher and only the wall, we have not the slightest difficulty in agreeing that the wall prevented the ball from hitting the window.

The real difference, I suggest, is that a wall is a more *stable* sort of thing than a catcher. Causal judgments seem to presuppose a division of events into (i) the "fixed background," and (ii) the "variables" (or "causal field") whose counterfactual relations must be considered. In the original case, the wall is naturally thought of as part of the fixed background of the story. The reasoning that leads us to revise our intuitive judgments forces us to think of the presence of the wall as a variable. In Collins's modified case, it is natural from the start to think of the second catcher's catching or noncatching as a relevant variable: since we are wondering what would have happened if the first catcher had missed, it is natural to wonder also what would have happened if the second catcher had missed.

---

[13] "Preemptive Prevention," this JOURNAL, XCVII, 4 (April 2000): 223-34.

Similarly in "The Switch," the reasoning that brings us to concede that the switching was after all a cause of the train's arrival seems to do so by forcing us to consider the presence of the unused track as a variable, rather than as part of the fixed background.

If this is correct, there is no ground for thinking that it forces us to correct a conceptual *mistake*. We give up our previous '*C* did not cause *E*', but *not* because it is shown to be inconsistent with '*C* caused *D* and *D* caused *E*'. (Analogy: suppose a man asserts, "Nothing is better than beer, and there is nothing in this glass." You can get him to withdraw the second conjunct by pointing out that there is surely some dust in the glass. But this does not show that his original assertion was *inconsistent*.)

Even if I am wrong about all this, and the facts of "The Switch" on the basis of which we judge that *C* caused *D* and *D* caused *E* really do also entail that *C* caused *E*, there is a further question. Does the same hold true for *all* apparent counterexamples to transitivity?

The answer, in my opinion, is "No." Consider "The Dog Bite": I press the button with my left forefinger. This was caused by the dog's biting off my right forefinger, and causes an explosion. But intuition says that the dog bite was not a cause of the explosion. Can this verdict be reversed by reasoning of the kind we have been considering? We need an actual event *X* of which we can say, "If *X* had not occurred, and the dog bite had not occurred, the explosion would not have occurred; so *X* and the dog bite between them caused the explosion; but *X* clearly contributed nothing." As far as I can see, there is no such event *X*. No such event, at any rate, is implied by the facts of the case, on the basis of which we judge that the dog bite caused the pressing with the left forefinger and that the pressing with the left forefinger caused the explosion.

Or, for a case with a quite different structure, consider "The Order": I order my two servants to push in opposite directions against a moveable object; it stays still. Clearly, the order caused servant *A* to push, and servant *A*'s push caused the object to stay still—it would have moved if *B*'s push had not been opposed. But intuition denies that the order caused the object to stay still. Can this verdict be reversed? As far as I can see, there is nothing to fill the role of event *X*—there is no actual event which actually "contributed nothing," but which the object's staying still *would* have been counterfactually dependent on if I had not given the order.

(The ultimate effect in this case was a nonevent, a staying still.[14] Some of my opponents would concede that transitivity can fail in such cases. But it can easily be modified to one in which E is a real event by any standards: add a third force who pushes at the crucial moment, so that the object moves; servant B's push is equal and opposite, and servant A's equal and in the same direction. Here the order causes servant A to push and servant A's push causes the object to move (it would have stayed still if he had not pushed); but the order does not cause it to move. The attempt to reverse this verdict still fails for want of a suitable X.)

I turn now to my alternative analysis of causation.[15] It is based on three more-or-less independent ideas.

VII. DIRECT CAUSATION

I favor an analysis of direct causation in terms of a notion of counterfactual sufficiency. A set of actual events C is sufficient for the occurrence of E if and only if E would have happened given C, *whatever else had happened.* For example, suppose that our relevant variables are represented by actual events $C_1, \dots, C_5, E$; then $\{C_1, C_2, C_3\}$ was sufficient for E if and only if the following four counterfactuals are all true:

> If C_1 and C_2 and C_3 had all occurred and C_4 (had/had not) occurred and C_5 (had/had not) occurred, E would have occurred.

Sufficiency is not transitive. The falling of the first domino is sufficient for the falling of the second, and the second for the third, but the falling of the first is not sufficient for the falling of the third. For if the first had fallen and the second *not* fallen, the third would not have fallen either. If the first had fallen the second *would* have fallen too, but (as Lewis has often pointed out) we can still sensibly ask what would have happened if it had not.

We can speak, more generally, of a sufficient *condition* for an event. To each actual event there corresponds a set of possible events that are its alternatives—the things that might otherwise have happened at that point. A *sufficient condition* for E is a condition on what happens at a point, or a conjunction of such conditions, such that given its satisfaction E would have occurred whatever had happened at *other* points. A *minimal* sufficient condition for E is a sufficient condition in which no conjunct could be replaced by a weaker condition on what happens at that point without losing sufficiency.

---

[14] I do not mean to concede that there is a serious distinction between events and omissions. Suppose I push a moveable object in a railway carriage, so that it moves relative to the carriage; simultaneously, the carriage moves, so that the object remains at rest relative to the ground. Is this real causation, with a real effect, a moving? Or is it just "causation" in some inferior sense, because the apparent "effect" is just an omission, a staying still?

[15] "Redundant Causation," *British Journal for the Philosophy of Science*, XL (1995): 523-44.

When C and E are distinct actual events, C is a *direct cause* of E, I suggest, if and only if the occurrence of C satisfies a conjunct in some satisfied minimal sufficient condition for E.

This sufficiency analysis shares the advantages of Lewis's new theory (or of the 'might' analysis) over the old theory. In "Two Buttons," for example, a minimal sufficient condition for the shock was that I should press a button. My pressing button A satisfied that condition, and hence was a direct cause of the shock. (That is on the simplest possible view of the case, where there are just two relevant variables—my pressing A (and what I could have done instead), and the shock (and what might have happened instead). We might wish to acknowledge further variables, such as the existence/construction/...of the connections between the buttons and victim—call this C^*. On this view, my pressing a button was not a sufficient condition for the shock. But its conjunction with C^* was, and a minimal condition to boot; my pressing button A satisfied the first conjunct, the occurrence of C^* satisfied the other; so my pressing A was still a cause of the shock. Or we might wish to notice events intermediate between button pressing and shock, in which case it would no longer be a case of *direct* causation.)

In the "Trumping" case, Merlin's frog spell was a minimal sufficient condition for the prince's frogification. No weaker condition on his action was: it is not true that, if he had done something else the frogification would still have occurred, *whatever else had happened*. So Merlin's spell was a cause.

As far as Morgana's spell is concerned, however, the sufficiency analysis seems to do better than the new theory as well. Morgana's spell *did* satisfy a conjunct in a satisfied minimal sufficient condition for the frogification, namely, the condition 'Morgana casts a frog spell and Merlin does *not* cast a goat spell'. As I said earlier, I think this is the right answer—Morgana's spell was also a cause.

The sufficiency analysis also solves the other problems of direct causation which baffled Lewis's new theory. In "The Neurons," C_1 and C_2 both count as causes of the firing of E. The firing of C_1, for example, was a minimal sufficient condition for the firing of E.

In the "Direct Early Cutting" case, Merlin's frog spell was a minimal sufficient condition for, and hence a cause of, the prince's frogification. The preempted cause, Morgana's preparedness, was not a sufficient condition for the frogification. For, given Morgana's preparedness, there would have been no frogification if neither Merlin nor Morgana had cast any spell.

The sufficiency analysis explains why causal judgments can be reversed by changing our assumptions about the causal field. Consider "The Catcher": the natural view is that the causal field is represented by the following actual events: the catcher's catching (*C*) and the ball's not hitting the window (not-*H*). On this view, *C* is a sufficient condition for not-*H*, but not a minimal sufficient condition: the null condition is also sufficient—not-*H* was bound to occur. So we say that *C* is *not* a cause of not-*H*. Now enlarge the causal field to include the presence/construction/...of the wall (*X*). *C* is still a sufficient condition for not-*H*: if *C* had occurred and *X* (had/had not) occurred, not-*H* would have occurred. But now *C* is a minimal sufficient condition. The null condition is no longer sufficient, since if *C* had not occurred and *X* had not occurred, not-*H* would not have occurred. That is why we *now* say that *C* caused not-*H*.

VIII. INDIRECT CAUSATION

The simplest way to turn the sufficiency analysis of direct causation into an account of causation in general would be to take the ancestral. That would guarantee the transitivity of causation. But I think we can do better.

Consider the following exchange:

> Peter: *C* caused *D* and *D* caused *E*. I conclude that *C* caused *E*.
>
> Paul: No, *C* was irrelevant—if *C* had not occurred E would still have occurred.
>
> Peter: Yes, but it would not have occurred *that way*.

If Peter's final remark is true, it seems to count in favor of his claim that *C* caused *E*. But how exactly should it be understood? What is a "way"?

A first thought is to individuate ways by the intermediate events involved: in the absence of *C*, there would have been a causal chain from not-*C* to *E*, but it would not have gone via *D*. This is the thinking which leads people to embrace transitivity,

My alternative suggestion is that ways should be individuated by their *sustaining conditions*—the minimal sufficient conditions, at points along the way, whose satisfaction directly contributed to the chain. What condition combined with *C* to cause *D*, and what condition combined with *D* to cause *E*? If *C* had not occurred, *E* would still have occurred, via different intermediate events. But the crucial question is, would different sustaining conditions have come into play? Or would the same sustaining conditions that led from *C* to *E* also have led from not-*C* to *E*? If they would, I say that *E* would have occurred in the *same* way without C as it did with C, and hence that C was *not* a cause of E.

In "The Dog Bite," for example, the bite did not cause my pressing with the left forefinger all by itself. The obvious sustaining condition was my desire for the button to be depressed. (The full minimal sufficient condition for the pressing with the left forefinger was the conjunction 'dog bite *and desire*'.) And the pressing of the button with the left forefinger did not cause the explosion all by itself. The obvious sustaining condition was the connection between the button and the bomb. The desire and the button-bomb connection constitute a sort of "channel." If you put in a dog bite at one end, you get an explosion out at the other end. But also if you put an absence of dog bite in at one end, you get the same explosion out at the other: the absence combines with the same desire as in actuality to produce a different intermediate event, which then combines with the same button-bomb connection as in actuality to produce the explosion. Even when we focus on the actual channel by which influence is transmitted, the dog bite was not necessary for the production of the explosion via that channel.

Contrast "The Switch": the switching caused the train to pass over the mid-point of the right-hand track; the obvious sustaining condition was the existence of the first half of the right-hand track. The train's passing over the mid-point of the right-hand track caused it to arrive at its destination; the obvious sustaining condition was the existence of the second half of the right-hand track. The right-hand track, then, is the channel. Put in the switching and you get out the train's arrival. But in this case the same channel would *not* yield the same result if you put in a *not*-switching by the engineer. If the engineer had not switched, the *left*-hand track, which was no part of the actually used channel, would have become crucial. The switching was essential in the production of the arrival via the actual channel.

In general, I suggest, C is a cause of E if and only if

(i) there is a chain of direct causation linking C to E; and
(ii) C is essential in the production of E *via that channel*: that is, either
(iia) without C, E might not have occurred, or
(iib) without C, E would have occurred via *another* channel—its occurrence would have been dependent on some satisfied condition X which "contributed nothing" to the actual chain of direct causation linking C to E.

IX. NONEXTENSIONALITY

The key to the problem of late cutting is that 'cause' does not stand for an extensional relation. Whether a sentence of the form 'C caused E' is true depends on how the event E is referred to. Suppose Frank serves badly, because of tension. We may assume that 'Frank's serve' and 'Frank's bad serve' refer to the same event. But 'Tension caused Frank's bad serve' is true and 'Tension caused Frank's serve' is false.

(The failure of extensionality in 'C caused E' seems to be *only* at the effect position. If Frank's bad serve dismayed his supporters, we shall happily assent also to 'Frank's serve dismayed his supporters [in virtue of its badness]'.)

When 'C caused E_1' is true and $E_1 = E_2$, we may not be ready to accept 'C caused E_2', but we are generally ready to accept '*C affected* E_2' instead. The relation of causing-or-affecting is, I think, extensional.

Let us assume extreme fragility (for the purposes of the analysis of causation). On that assumption, our current analysis will do nicely as an account of the extensional relation of causing-or-affecting. In particular, we have the first part of a solution to the problem of late cutting. Suzy's throw was linked to the bottle's shattering by a chain of direct causation, and without her throw the shattering, in all its precise detail as to time and manner of occurrence, would not have occurred. So 'Suzy's throw caused or affected the shattering' is true. Billy's throw is linked to the shattering (the actual, extremely fragile, shattering) by no such chain. So Billy's throw neither caused nor affected the shattering.

To separate causing from affecting, we need one further clause in our analysis: 'C caused E' is true if and only if

(i) there is a chain of direct causation linking C to E; and
(ii) C is essential in the production of E, via that channel; and
(iii) C is essential in the production of an event meriting the description 'E', via that channel.

This solves the problem of spurious causation. In "The Doctor," for example, there is a chain of direct causation linking his treatment with the actual, extremely fragile, death, and his treatment was essential in the production of that death via that channel. But the treatment was not essential in the production of *a* death via that channel: given the other sustaining conditions of the actual process, the patient would have died, sooner or later, irrespective of whether the doctor had treated the patient or not (and independently of anything outside the actual process). So 'The doctor's treatment caused the patient's death' is false.

'Suzy's throw caused the bottle's shattering', however, is true. Her throw was essential in the production of *a* shattering via the actual channel. Certainly, there would have been a shattering if she had not thrown; but not via the channel which actually led to the actual (fragile) shattering: if Suzy had not thrown, the occurrence of a shattering would have been dependent on Billy's throw, which was not a sustaining condition of the actual channel.

Since our analysis does nothing to modify the effect of the extreme-fragility assumption as far as the *cause* is concerned, it gives the right answer in "The Concert." The concert was a sufficient condition for the riot, and a minimal one because if it had not occurred—that is, if it had not occurred exactly as it did—the riot might not have occurred.

MICHAEL MCDERMOTT

University of Sydney

BOOK REVIEWS

On Clear and Confused Ideas: An Essay about Substance Concepts. RUTH GARRETT MILLIKAN. New York: Cambridge University Press, 2000. xiii + 258 p. Cloth $59.95.

Ruth Garrett Millikan's book resists easy summary. Indeed, we suspect we may have a confused idea concerning what this book is about. In what follows, we attempt to convey our conception of the main themes. Our primary concern is to canvass the contributions rather than to criticize the arguments. Although our treatment is necessarily selective, we hope it is representative.

The engine that drives this book is a broad theory of the nature of substance concepts as abilities to identify "things" that are nexii for inductive inference:

> Substances...are whatever one can learn from given only one or a few encounters, various skills or information that will apply on other encounters. Further, this possibility must be grounded in some kind of natural necessity. The function of a substance concept is to make possible this sort of learning and use of knowledge for a specific substance (33).

Substances, then, need not be individuals; real kinds (flamingos, ice dancing) and stuffs (water, dirt) are fine, too, since one can learn things about each of these with some confidence that they (and their implicated kin) will apply on other occasions, or to other samples.[1] The central thesis of the book is that a substance concept is an ability to identify that substance.

Millikan has a good deal to say about abilities. For present purposes, the important point is this: having an ability to identify *X* is just to harbor something or other—call it an *identifier*—whose function it is to identify *X*. Millikan states: "Substance concepts are abilities of a certain kind. They are, in part, abilities to reidentify their assigned substances" (51). Abilities are not dispositions, but imply them. To have an ability in her sense is to have some disposition to succeed at identifying a substance in normal conditions, where normal conditions are "conditions under which that function was performed historically such as to be selected for" (67). In other words, the

---

[1] This idea should be familiar from Millikan's "How We Make Our Ideas Clear," *Proceedings and Addresses of the American Philosophical Association*, LXXII, 2 (November 1998): 65-79.

0022-362X/02/9902/102–08

extension of a substance concept is *X* if tracking/identifying *X*s in normal circumstances is what accounts for selection of that tracking/identifying mechanism.[2]

Conceptions—knowledge/belief about the target substance—are among the means one has at one's disposal to identify that substance, but they do not determine the extensions of the substance concepts they service. One's water concept must be kept distinct from one's conception of water. For one thing, you and I can have arbitrarily different conceptions of water, yet share an ability to identify the same stuff, hence share a concept of water. Moreover, my conception of water need not be very good: I can have all sorts of misconceptions about it, provided that they are part of a mechanism or process whose function it is to identify water (and not *XYZ* or alcohol). Conceptions can be "built up" out of perceptual mechanisms that allow for perceptual tracking, and conceptual mechanisms that allow for conceptual tracking—that is, understanding, instead of perceiving, that something is the same.

Putting these points together, we have the basis for a kind of realism *cum* direct reference theory. The extension of a concept *A* is simply whatever it is that *A* is an ability to identify. As we have just seen, *A* can be an ability to identify something about which one has a good deal of misinformation. Two people, or the same person at different times, can be said to share a concept even though their conceptions differ significantly.

Acquisition and deployment of substance concepts—that is, building the relevant conceptions—is facilitated by what Millikan calls *substance templates*. These are bodies of information about sorts of things: for example, animals, persons, and chemical compounds. When you learn that an aardvark is an animal, you learn the sorts of properties that are likely to remain stable in an aardvark over time: behavior, physiology, appearance, and so on. "You can ask how tall Mama is, but not how tall gold is. You can ask at what temperature gold melts, but not at what temperature chairs (as such) do—the latter is a question that can be answered only for certain individual chairs" (9).

One of the more interesting ideas in the book is that language is (among other things) a perceptual modality, a "direct medium for the perception of objects" (84). Evidently, many of our substance concepts are not acquired through encounters with the substances themselves, but only by hearing or reading about them. Millikan argues that hearing about (or reading about) an aardvark is a way of perceiving one. Just two things, Millikan claims, distinguish acquiring information about aardvarks via direct perception from acquiring information about aardvarks via linguistic description. Sensory perception typically specifies the perceiver's spatial relation to the object perceived, and is a great deal more reliable than verbal description.

---

[2] The selection in question here is, of course, not typically natural selection (since most substance concepts are learned) but the sort of functions Millikan thinks are determined by what she calls *adapted* and *derived proper functions*. These are explained in Millikan's *Language, Thought and Other Biological Categories* (Cambridge: MIT, 1984), chapter 2.

> Recognizing a linguistic reference to a substance is just another way of reidentifying the substance itself. It is identifying it through one more medium of manifestation. Think of this medium as like an instrument that aids perception. Like a camera, a radio, a cat scan, or a microscope, another person who talks to me picks up information-bearing patterns from his environment, focuses them, translates them into a new medium and beams them to me. Or think of living in a language community as like being inundated in one more sea of ambient energy. Like the surrounding light, surrounding people transmit the structure of the environment to me in ways that, barring certain interferences, I can become tuned to interpret (89).

The abilities that support substance concepts are abilities to *re*identify their objects correctly. Millikan prefers to use 'reidentify' or 'coidentify' in this context in order to emphasize the fact that every act of identification involves more than one thought.[3] Correct identification of a substance is achieved via the *co*identification of those representational vehicles which bear information about it: "In an act of coidentifying, two representational vehicles are employed together in a manner that assumes, that is, requires for correctness, an overlap or partial identity in content, thus effecting an act of reidentifying of content" (144).

Millikan claims that the received view is that reidentifying an object (or property) consists in the system's "repeating some kind of response" (109). She calls this the *repetition view of reidentifying* (RVR). The error that gives rise to RVR is that of confusing properties of the vehicles of representation with properties of their representeds, or "intentional contents." In particular, RVR results when the property of sameness, or likeness, in vehicle is taken to be a representation *of* sameness. Millikan argues that repetition of response—for example,

---

[3] For an extended discussion of coidentification, see Millikan's "Knowing What I'm Thinking Of," *Proceedings of the Aristotelian Society*, Supplementary Volume LXVII (1993): 109-24.

"attaching" the same mental term or thought to something—does not constitute an act of reidentification. Perceptual tracking of substances, for example, can proceed by means of the system reidentifying the substance solely by its perceived *continuity* in space and time. In such cases, it is not the *duplication* of "percepts" that is taken as a mark of sameness:

> Given the right continuities, one's perception may be of an object as being the same one over a period of perceptual tracking despite its apparently changing in every one of its observed properties. This sort of tracking of an object, say, with the eyes, does not involve *repeating* some particular way of perceiving or thinking of the object, or repeating some way of recognizing it over and over. It is not repetition that constitutes or triggers perceptual grasp of identity (139).

In order to determine in what the act of reidentifying actually consists, Millikan claims we must first determine the purpose or function of identifying. She takes the reidentification of substances to serve the function of enabling the system to track a substance over time, so that it may gather information about the same *as* about the same, which facilitates action, learning, and theoretical and practical inference. Recognizing sameness is then taken to consist, first, in the system's *marking* its representations as about the same. Millikan calls this a *first act identification*, which counts as a first act of grasping sameness. Once marked as about the same, the representations are ready to be used together *as* coidentified: "a 'second act understanding' is an actual process of mediate inference, amplificatory information processing, or action guidance controlled by these markers. First act identifications, sameness markings, prepare for second act identifications" (144). Exactly which features of a system's representations it uses to mark sameness of reference and just how it goes about marking identity, are matters for empirical investigation.

Millikan has a lot to say about Fregean senses (and their graspings), and why a representational theory of mind must abandon them. There is simply not the space to detail each of her objections. Suffice it to say, as she does herself, that Gottlob Frege did not take senses to be representations anyway, but abstract objects. More pertinent to the current scene are her criticisms of the *language of thought* (LOT).

RVR is operative in LOT to the extent that the act of reidentification is taken to consist in the reoccurrence, or application, of the same mental term. As we have seen, Millikan takes RVR to be false. But it is, surprisingly, Millikan's charge that LOT is incompatible with *externalism* that constitutes her most novel critique of the model:

> I have been arguing in this book for an externalist and representationalist position on thought, but it is crucial that we not rely on a language-of-thought model of mental representation. This misleading model offers an image of thought as having two levels of sameness marking, on one of which sameness of content for concepts is not an empirical issue but a matter of mere inner form. It pictures errors about identity as impossible on this level, and, correlatively, pictures the difference between valid and invalid mediate inference as though it were distinctly marked from within. But on externalist grounds, no distinction can be clearly marked in thought between valid inference that relies on false identity premises and invalid inference. There is no analytic/synthetic distinction for identity as grasped in thought (171).

On Millikan's view, correct grasp of identity requires the cooperation of the distal environment. The fact that the system coidentifies representations as of the same is no guarantee that their reference *is* the same:

> Putting the inevitably a posteriori nature of our grasp of sameness in John Campbell's terms (1987/88),[4] there is no such thing as completely "manifest sameness of reference." The mental sentence image causes us to overlook the most central fact about cognition, namely, that its most difficult job is to get the empirical identities right, to create a coherent, nonredundant and nonequivocal mental representational system (171-72).

Even if RVR were true, it would be inconsistent with externalism to hold that sameness of inner form is an unmistakable mark of identity of reference. Every system must be considered in principle *fallible* at reidentifying, and thus there can be no a priori, certain, *inner* mark of sameness of content. Millikan concludes that we must reject the LOT model as incompatible with a thoroughgoing externalism about mental content.

While Millikan's primary aim in this book is to develop a view about the nature and structure of substance concepts, she also addresses epistemological worries that meaning externalist theories of content generate. She argues that, for a meaning externalist position like her own, there is no a priori guarantee against possessing empty, equivocal, or duplicate concepts; only through experience can we determine if our concepts are nonempty, unequivocal, and nonduplicating. A concept is likely to be objective (nonempty) and unequivocal if alternate tests converge to the same judgment. For example, the concept *temperature forty degrees centigrade* is objective since independent methods for measuring the temperature of some object *X* yield the same judgment: *X* has a temperature of forty degrees centigrade. That independent methods yield the same judgment is evidence that the concept is objective, regardless of the truth of our theories of temperature, and regardless of the truth of our theories of how thermometers work. She puts this point by saying that "[t]he epistemology of concepts is prior to and not the same as the epistemology of judgment" (102).

---

[4] The reference is to Campbell, "Is Sense Transparent?" *Aristotelian Society Proceedings*, New Series, Volume LXXXVIII (1987/1988): 273-92.

When applied to nontheoretical terms, our ability to come to the same judgment using disparate sensory modalities confirms the objectivity of the concepts employed in the judgment. Recognizing contradiction in judgment also reveals equivocation in concepts, so in order to test concepts empirically one must be able to recognize complements or contraries of predicates. Millikan writes: "Not observing that the apple is red does not equal observing that it is not red. To tell that it is not red I must be able to tell what other color it is instead, that it is some contrary of red or, more generally, that it is non-red, the complement of red" (106). Finally, redundancy of concepts shows up "with the accumulation of coincident properties and the absence of contrary ones" (103).

Millikan is also concerned to show what knowing what one is thinking about amounts to on an externalist account of meaning. She rejects the idea that knowing what one is thinking about is propositional knowledge or knowledge *that*, as well as the Russellian view that knowledge of what one is thinking about consists in acquaintance with the objects of one's thought, where the object is taken to be literally in the mind. Rather, she views this knowledge as knowledge *how*, or an ability. This ability, as it turns out, is just the ability to re-identify the object. In place of the confrontation of thought with object, we get the "confrontation of one *thought* with another *thought* concerning the same object, this taking place within thought itself, and constituting a recognition *of* the sameness of the object.... Knowing what I am thinking of is being capable of coidentifying various of my thoughts with other thoughts of the same" (184). So knowing what one is thinking of is simply having the ability to coidentify thoughts of the object of one's thoughts with thoughts of the same object.

There are various topics covered in some detail in this book which we have not even mentioned. Most notably, we have not discussed Millikan's treatments of various positions in the literature. In particular, P. F. Strawson and Gareth Evans are treated at some length (chapter 11 and passim). We have bypassed an interesting discussion of laws in psychology (chapter 15), and a discussion of conceptual and nonconceptual representation (chapter 14). We have tried instead to convey a sense of what is theoretically novel in the book. As is always the case with Millikan's work, however, the details are worth studying with one of the philosophy of mind's most original thinkers.

ROBERT CUMMINS
ALEXA LEE
MARTIN ROTH
DAVID BYRD
PIERRE POIRIER

University of California/Davis

NEW BOOKS: TRANSLATIONS

ADORNO, THEODOR W.: *Metaphysics: Concept and Problems.* ROLF TIEDEMANN, ed., EDMUND JEPHCOTT, trans. Stanford: University Press, 2000. ix, 214 p. Cloth $39.50.

ARISTOTLE: *Aristotle's On the Soul and On Memory and Recollection.* JOE SACHS, trans. Santa Fe: Green Lion, 2001. x, 211 p. Cloth $32.00, paper $19.95.

AUGUSTINE: *Political Writings.* E.M. ATKINS and R.J. DODARO, eds., MARGARET ATKINS, trans. New York: Cambridge, 2001. li, 299 p. Cloth $64.95, paper $22.95.

BATAILLE, GEORGES: *The Unfinished System of Nonknowledge.* STUART KENDALL and MICHELLE KENDALL, trans. Minneapolis: Minnesota UP, 2001. xliv, 305 p. Cloth $39.95.

BAUDRILLARD, JEAN: *Jean Baudrillard: Selected Writings,* Second Edition, Revised and Expanded. MARK POSTER, ed., various trans. Stanford: University Press, 2001. vii, 294 p. Cloth $49.50, paper $17.95.

BESANÇON, ALAIN: *The Forbidden Image: An Intellectual History of Iconoclasm.* JANE MARIE TODD, trans. Chicago: University Press, 2000. viii, 423 p. Cloth $40.00.

BÖHME, GERNOT: *Ethics in Context: The Art of Dealing with Serious Questions.* EDMUND JEPHCOTT, trans. Malden, MA: Blackwell, 2001. 189 p. Cloth $56.95, paper $21.95.

BUKDAHL, JØRGEN: *Søren Kierkegaard and the Common Man.* BRUCE KIRMMSE, trans. and ed. Grand Rapids, MI: Eerdmans, 2001. xviii, 154 p. Cloth $19.00.

CONDILLAC, ETIENNE BONNOT DE: *Essay on the Origin of Human Knowledge.* HANS AARSLEFF, trans. and ed. New York: Cambridge, 2001. xlv, 225 p. Cloth $59.95, paper $21.95.

DELEUZE, GILLES: *Pure Immanence: Essays on a Life.* ANNE BOYMAN, trans. Cambridge: MIT, 2001. 102 p. Cloth $24.00.

DERRIDA, JACQUES: *On Cosmopolitanism and Forgiveness.* MARK DOOLEY and MICHAEL HUGHES, trans. New York: Routledge, 2001. xii, 60 p. Cloth $50.00, paper $12.95; *The Work of Mourning.* PASCALE-ANNE BRAULT and MICHAEL NAAS, eds., various trans. Chicago: University Press, 2001. viii, 262 p. Cloth $25.00.

DERRIDA, JACQUES, and MAURIZIO FERRARIS: *A Taste for the Secret.* GIACOMO DONIS, trans., GIACOMO DONIS and DAVID WEBB, eds. Malden, MA: Blackwell, 2001. viii, 161 p. Cloth $52.95, paper $19.95.

DESCARTES, RENÉ: *Discourse on Method, Optics, Geometry, and Meteorology,* Revised Edition. PAUL J. OLSCAMP, trans. Indianapolis: Hackett, 2001. xxxvi, 388 p. Cloth $45.00, paper $16.95.

DYENS, OLLIVIER: *Metal and Flesh: The Evolution of Man: Technology Takes Over.* EVAN J. BIBBEE and OLLIVIER DYENS, trans. Cambridge: MIT, 2001. 134 p. Cloth $24.95.

ERASMUS: *The Adages of Erasmus.* WILLIAM BARKER, ed., various trans. Toronto: University Press, 2001. li, 405 p. Cloth $80.00, paper $29.95.

FLYVBJERG, BENT: *Making Social Science Matter: Why Social Inquiry Fails and How It Can Succeed Again.* STEVEN SAMPSON, trans. New York: Cambridge, 2001. x, 204 p. Cloth $54.95, paper $19.95.

GINZBURG, CARLO: *Wooden Eyes: Nine Reflections on Distance.* MARTIN RYLE and KATE SOPER, trans. New York: Columbia, 2001. xv, 261 p. Cloth $24.95.

HEIDEGGER, MARTIN: *Phenomenological Interpretations of Aristotle.* RICHARD ROJCEWICZ, trans. Bloomington: Indiana UP, 2001. xiv, 160 p. Cloth $37.95; *The Fundamental Concepts of Metaphysics: World, Finitude, Solitude.* WILLIAM MCNEILL and NICHOLAS WALKER, trans. Bloomington: Indiana UP, 2001. xxi, 376 p. Cloth $39.95, paper $19.95; *Zollikon Seminars: Protocols, Conversations, Letters.* FRANZ MAYR and RICHARD ASKAY, trans., MEDARD BOSS, ed. Evanston, IL: Northwestern, 2001. xxxiii, 360 p. Cloth $89.50, paper $24.95.

HUSSERL, EDMUND: *The Shorter Logical Investigations.* J.N. FINDLAY, trans., DERMOT MORAN, ed. New York: Routledge, 2001. lxxxv, 422 p. Paper $27.95.

IRIGARAY, LUCE: *Democracy Begins between Two.* KIRSTEEN ANDERSON, trans. New York: Routledge, 2001. xi, 224 p. Cloth $70.00, paper $18.95; *To Be Two.* MONIQUE M. RHODES and MARCO F. COCITO-MONOC, trans. New York: Routledge, 2001. vii, 125 p. Cloth $80.00, paper $18.95.

JOAS, HANS: *The Genesis of Values.* GREGORY MOORE, trans. Chicago: University Press, 2000. xiv, 250 p. Cloth $50.00, paper $22.50.

LOSURDO, DOMENICO: *Heidegger and the Ideology of War: Community, Death, and the West.* MARELLA MORRIS and JOHN MORRIS, trans. Amherst, NY: Humanity, 2001. 256 p. Cloth $50.00.

LYOTARD, JEAN-FRANÇOIS: *Soundproof Room: Malraux's Anti-Aesthetics.* ROBERT HARVEY, trans. Stanford: University Press, 2001. xi, 116 p. Cloth $39.50, paper $12.95.

NANCY, JEAN-LUC: *The Speculative Remark (One of Hegel's Bons Mots).* CÉLINE SURPRENANT, trans. Stanford: University Press, 2001. xx, 196 p. Cloth $49.95, paper $16.95.

POPPER, KARL: *All Life Is Problem Solving.* PATRICK CAMILLER, trans. New York: Routledge, 2001. xii, 171 p. Paper $16.95.

TODOROV, TZVETAN: *Frail Happiness: An Essay on Rousseau.* JOHN T. SCOTT and ROBERT D. ZARETSKY, trans. University Park: Penn State UP, 2001. xxxii, 70 p. Cloth $22.50; *Life in Common: An Essay in General Anthropology.* KATHERINE GOLSAN and LUCY GOLSAN, trans. Lincoln: Nebraska UP, 2001. xii, 174 p. Cloth $45.00, paper $19.95.

WALTER, HENRIK: *Neurophilosophy of Free Will: From Libertarian Illusions to a Concept of Natural Autonomy.* CYNTHIA KLOHR, trans. Cambridge: MIT, 2001. xiii, 391 p. Cloth $45.00.

XENOPHON: *The Education of Cyrus.* WAYNE AMBLER, trans. Ithaca: Cornell, 2001. x, 304 p. Cloth $45.00, paper $19.95.

THE JOURNAL OF
PHILOSOPHY

The Journal is pleased to announce to our readers and subscribers the launch of its new website, which will have the pages listed below. The address is: *www.journalofphilosophy.org* or simply *www.jphil.org.*

Purpose & History
Contacts and Addresses
Board of Trustees
Editorial Board
Memos to Authors
Subscriptions & Advertising

Recent & Forthcoming Issues
Electronic Issues
Index 1904–present
Letters to the Editor
Rights and Permissions
Lectures & Conferences
Philosophy Links

* * *

ue to an administrative error by Routledge made in connection ith the preparation of the 'Predicate Calculus' entry in the outledge Encyclopedia of Philosophy, the entry that appeared ove Professor Lavine's name contains various statements that ofessor Lavine did not authorize and with which he does not ree.

* * *

Classic Works in the History of Ideas from Thoemmes Press

Selected Works of James Mark Baldwin

Edited by ROBERT WOZNIAK

2900 pages, 7 Volumes 1-85506-916-4
Cloth $595.00

The Edinburgh Philosophical Journal

Edited by DAVID BREWSTER and ROBERT JAMESON

6400 pages, 14 Volumes 1-85506-923-7
Cloth $2240 00

Life and Collected Works of Thomas Brown

Edited by THOMAS DIXON

4700 pages, 8 Volumes 1-85506-924-5
Cloth $895.00

Aesthetics and the Picturesque, 1795–1840

Edited by GAVIN BUDGE

2500 pages, 6 Volumes 1-85506-900-8
Cloth $750.00

For more information about these and other Thoemmes Press publications, visit the Thoemmes Press Web site at www.thoemmes.com.

Lectures, Sermons and Dissertations

GEORGE CAMPBELL

Edited by Doug Sonheim

1200 pages, 3 Volumes 1-85506-902-4
Cloth $375.00

An Essay on Philosophical Necessity

ALEXANDER CROMBIE

544 pages 1-85506-003-5
Cloth $135.00

American New Realism 1910-1920

Edited by CORNELIS DE WAAL

750 pages, 3 Volumes 1-85506-860-5
Cloth $295.00

Philosophical Works of James Frederick Ferrier

Edited by A. GRANT and E. L. LUSHINGTON

1770 pages, 3 Volumes 1-85506-835-4
Cloth $360.00

Early Responses to Hume's Writings on Religion

Edited by JAMES FIESER

800 pages, 2 Volumes 1-85506-797-8
Cloth $280.00

Distributed for Thoemmes Press UK

The University of Chicago Press
1427 East 60th Street, Chicago, IL 60637 www.press.uchicago.edu

Classic Works in the History of Ideas from Thoemmes Press

Philosophical and Literary Essays, and Letters

JAMES GREGORY

Edited by Alexander Broadie and James Fieser

1520 pages, 3 Volumes 1-85506-879-6
Cloth $360.00

Works of William Hamilton

Edited by SAVINA TROPEA

2822 pages, 6 Volumes 1-85506-920-2
Cloth $750.00

The Reception of British Aesthetics in Germany

Seven Significant Translations, 1745–1776

Edited by HEINER F. KLEMME and MANFRED KUEHN

2100 pages, 7 Volumes 1-85506-899-0
Cloth $875.00

Analysis of the Phenomena of the Human Mind

JAMES MILL

648 pages, 2 Volumes 1-85506-918-0
Cloth $225.00

For more information about these and other Thoemmes Press publications, visit the Thoemmes Press Web site at www.thoemmes.com.

A Fragment on Mackintosh

JAMES MILL

444 pages 1-85506-919-9
Cloth $120.00

Essays and Observations, Physical and Literary

Royal Society of Edinburgh

1590 pages, 3 Volumes 1-85506-925-3
Cloth $445 00

Early Critics of Pragmatism

Edited by JOHN R. SHOOK

1680 pages, 5 Volumes 1-85506-914-8
Cloth $450.00

Einleitung in die Philosophie

JOHANN GEORG WALCH

818 pages, 2 Volumes 1-85506-896-6
Cloth $240.00

Philosophisches Lexicon

JOHANN GEORG WALCH

1720 pages, 3 Volumes 1-85506-895-8
Cloth $470.00

Distributed for Thoemmes Press UK

The University of Chicago Press
1427 East 60th Street, Chicago, IL 60637 www.press.uchicago edu

Best in Philosophy

Cicero: *On Moral Ends*
Marcus Tullius Cicero
Julia Annas, Editor
Cambridge Texts in the History of Philosophy
0-521-66061-0, Hardback, $38.00
0-521-66901-4, Paperback, $14.00

Condillac: *Essay on the Origin of Human Knowledge*
Etienne Bonnot de Condillac
Hans Aarsleff, Editor
Cambridge Texts in the History of Philosophy
0-521-58467-1, Hardback, $60.00
0-521-58576-7, Paperback, $22.00

Intellectual Trust in Oneself and Others
Richard Foley
Cambridge Studies in Philosophy
0-521-79308-4, Hardback, $55.00

Physicalism and Its Discontents
Carl Gillett and
Barry Loewer, Editors
0-521-80175-3, Hardback, $60.00

Practical Rules
When We Need Them and When We Don't
Alan H. Goldman
Cambridge Studies in Philosophy
0-521-80729-8, Hardback, $55.00

The Structure of Values and Norms
Sven Ove Hansson
Cambridge Studies in Probability, Induction and Decision Theory
0-521-79204-5, Hardback, $60.00

Divine Hiddenness
New Essays
Daniel Howard-Snyder and
Paul K. Moser, Editors
0-521-80353-5, Hardback, $60.00
0-521-00610-4, Paperback, $22.00

G.E. Moore's Ethical Theory
Resistance and Reconciliation
Brian Hutchinson
0-521-80055-2, Hardback, $55.00

Wittgenstein
Biography and Philosophy
James C. Klagge, Editor
0-521-80397-7, Hardback, $55.00
0-521-00868-9, Paperback, $20.00

Fichte
The Self and the Calling of Philosophy, 1762-1799
Anthony J. La Vopa
0-521-79145-6, Hardback, $55.00

Enlightenment and Action from Descartes to Kant
Passionate Thought
Michael Losonsky
0-521-80612-7, Hardback, $55.00

Evil and the Augustinian Tradition
Charles T. Mathewes
0-521-80715-8, Hardback, $60.00

Collective Rationality and Collective Reasoning
Christopher McMahon
0-521-80462-0, Hardback, $55.00
0-521-01178-7, Paperback, $20 00

Social Action
A Teleological Account
Seumas Miller
0-521-78316-X, Hardback, $55.00
0-521-78886-2, Hardback, $20.00

Practical Rationality and Preference
Essays for David Gauthier
Christopher W. Morris and
Arthur Ripstein, Editors
0-521-78184-1, Hardback, $55.00

from Cambridge

New Essays in the Legal and Political Theory of Property
Stephen R. Munzer, Editor
Cambridge Studies in Philosophy and Law
0-521-64001-6, Hardback, $55.00

Natural Law and Practical Rationality
Mark C. Murphy
Cambridge Studies in Philosophy and Law
0-521-80229-6, Hardback, $55.00

Nietzsche: *The Gay Science*
Friedrich Nietzsche
Bernard Williams, Editor
Cambridge Texts in the History of Philosophy
0-521-63159-9, Hardback, $40.00
0-521-63645-0, Paperback, $12.00

Philosophy at the New Millennium
Anthony O'Hear
Royal Institute of Philosophy Supplements 48
0-521-00508-6, Paperback, $25.00

Margaret Cavendish: *Observations upon Experimental Philosophy*
Eileen O'Neill
0-521-77204-4, Hardback, $60.00
0-521-77675-9, Hardback, $23.00

Virtue Ethics and Professional Roles
Justin Oakley and Dean Cocking
0-521-79305-X, Hardback, $60.00

Thomas Aquinas on Human Nature
A Philosophical Study of *Summa theologiae*
Robert Pasnau
0-521-80732-8, Hardback, $75.00
0-521-00189-7, Paperback, $28.00

Antiphon the Sophist: *The Fragments*
Jerry Pendrick
0-521-65161-1, Hardback, $75.00

Philosophy and the Law of Torts
Gerald J. Postema
0-521-62282-4, Hardback, $60.00

Value, Respect, and Attachment
Joseph Raz
The Seeley Lectures 4
0-521-80180-X, Hardback, $55.00
0-521-00022-X, Paperback, $19.00

The Cambridge Companion to Rousseau
Patrick Riley, Editor
Cambridge Companions to Philosophy
0-521-57265-7, Hardback, $70 00
0-521-57615-6, Paperback, $25.00

Richard Wollheim on the Art of Painting
Art as Representation and Expression
Rob van Gerwen, Editor
0-521-80174-5, Hardback, $55.00

Decision Space
Multidimensional Utility Analysis
Paul Weirich
Cambridge Studies in Probability, Induction and Decision Theory
0-521-80009-9, Hardback, $60.00

Sameness and Substance
Second Edition
David Wiggins
0-521-45411-5, Hardback, $50.00
0-521-45619-3, Paperback, $22 00

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS

800-872-7423
us.cambridge.org/philosophy

ROUTLEDGE

NEW IN PHILOSOPHY

ROUTLEDGE PHILOSOPHY GUIDEBOOK TO ARISTOTLE ON ETHICS

Gerard J. Hughes

Hughes explains the key elements in Aristotle's *Nichomachaean Ethics* and highlights the controversy regarding the interpretations of his writings. In addition, he examines the role that Aristotle's ethics continue to play in contemporary moral philosophy by comparing and contrasting his views with those widely held today.

Pb • $14.95

ROUTLEDGE PHILOSOPHY GUIDEBOOK TO HUME ON RELIGION

David O'Connor

Hume on Religion introduces his major work on the subject, *Dialogues Concerning Natural Religion* and assesses Hume's life and the background to the work. The ideas and text of the work are also considered along with Hume's continuing importance to philosophy today.

Pb • $14.95

PHILOSOPHY, LITERATURE AND THE HUMAN GOOD

Michael Weston

Weston provides an exciting and challenging new interpretation of the elements of literature which concern our reflections on, and choices about, what constitutes the good life. The author examines the work of such thinkers as Bataille, Derrida, Kierkegaard, Rorty, Nussbaum, Murdoch and Blanchot

Pb • $24.95

Introducing the Thinking in Action series

ON RELIGION
John D. Caputo

ON SCIENCE
B.K. Ridley

ON BELIEF
Slavoj Žižek

ON THE INTERNET
Hubert L. Dreyfus

ON FILM
Stephen Mulhall

ON STORIES
Richard Kearney

ON COSMOPOLITANISM AND FORGIVENESS
Jacques Derrida

Each volume $12.95

www.routledgeny.com

1.800.634.7064

ROUTLEDGE Taylor & Francis Group

THE JOURNAL OF PHILOSOPHY

28\5\02

VOLUME XCIX, NUMBER 3
MARCH 2002

page
111 *Why There Is No Issue between Habermas and Rawls* Christopher McMahon
130 *De Facto Dependence* Stephen Yablo

BOOK REVIEWS
149 Allen Wood: *Kant's Ethical Thought* Stephen Engstrom
153 Richard Wollheim: *On the Emotions* Sabina Lovibond
157 Peter Geach: *Truth and Hope* John Haldane
163 Richard Sorabji: *Emotion and Peace of Mind: From Stoic Agitation to Christian Temptation* Constance C. Meinwald

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.
You may also visit our website at www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2002

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 2002 by the Journal of Philosophy, Inc.

ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCIX, NO. 3, MARCH 2002

WHY THERE IS NO ISSUE BETWEEN HABERMAS AND RAWLS

In a recent exchange between Jürgen Habermas and John Rawls,[1] Habermas describes one important issue dividing the two of them as a disagreement about the proper interpretation of the moral point of view (*ibid.*, pp. 116-17). For Rawls, the moral point of view is constituted by the constraints associated with the original position. For Habermas, by contrast, adopting it involves a "joint process of 'ideal role taking'" in which everyone is required to take the perspective of everyone else. He says: "from this interlocking of perspectives there emerges an ideally extended we-perspective from which all can test in common whether they wish to make a controversial norm the basis of their shared practice" (*ibid.*, p. 117).[2]

I believe, however, that when the moral theories of Habermas and Rawls are examined more fully, it can be seen that these two procedures of moral thinking are not in competition with each other. Rawls can make use of Habermas's procedure, and Habermas could employ a theoretical device that plays the role of Rawls's original position. One further point of contrast is that Rawls's most recent work is focused on identifying the correct principles of justice for a modern pluralistic democracy, while Habermas intends to provide a characterization of the moral point of view suitable for all moral reasoning. Still, for Habermas, moral principles are principles of

---

[1] Habermas, "Reconciliation through the Public Use of Reason: Remarks on John Rawls's Political Liberalism," this JOURNAL, XCII, 3 (March 1995): 109-31; Rawls, "Reply to Habermas," this JOURNAL, XCII, 3 (March 1995): 132-80.

[2] Rawls seems to concede the point when he speaks of "differences between our *devices of representation*, as I call them: his [Habermas's] is the ideal discourse situation as part of his theory of communicative action and mine is the original position" (*op. cit.*, p. 132)

0022-362X/02/9903/111–29

justice.[3] So we can interpret Habermas and Rawls as advancing different proposals concerning how to understand the requirements of justice in political contexts.

I

Anyone familiar with Habermas's moral philosophy will be aware that he posits a distinction between moral theories of two fundamentally different kinds: *monological* and *dialogical*. Roughly, the distinction turns on the extent to which the identification of the correct principles of morality is a project that must be carried out collectively. Dialogical theories affirm this and monological theories—implicitly, at least—deny it.

Viewed in one way, however, this distinction has little bite. Consider Rawls's theory of justice. Habermasians regard this as a monological theory. But the method of Rawls's theory is the method of reflective equilibrium, which produces a characterization of the original position, and thus of the true principles of justice, by balancing against each other considered judgments and general theoretical considerations.[4] This method can, and probably should, be carried out collectively. That is, it is entirely in order for individuals employing the method of reflective equilibrium to exchange information and arguments germane to the construction of the original position and the determination of what would be chosen there. What, then, do Habermasians mean by calling Rawls's theory monological?

The key appears to be the way the original position functions. Although Rawls speaks as if a number of parties occupy the original position, the procedure by which a conclusion concerning the correct principles of justice is obtained is one that can be followed by a single, representative individual. It is simply a matter of determining which principle would yield the best worst position, where this is measured by an index of primary goods which is assumed to be the same for everyone.[5]

We can generalize this observation using the concept of the *moral point of view*. The moral point of view is the standpoint from which the true principles of morality are revealed. Rawls's original position can be understood as an interpretation of the moral point of view. This interpretation is established by the method of reflective equilibrium, and those utilizing this method can reason collectively. But as Rawls characterizes the moral point of view, operating within it need not involve any kind of collective activity. It is for this reason that his theory counts as monological.

---

[3] See, for example, Habermas's *Between Facts and Norms*, William Rehg, trans. (Cambridge: MIT, 1996), p 153. Habermas distinguishes between morality and ethics. Ethics is grounded in culturally specific values, while morality is grounded in universal principles of justice.

[4] For an account, see Norman Daniels, *Justice and Justification* (New York: Cambridge, 1996).

[5] As Rawls puts it: "Habermas sometimes says that the original position is monological and not dialogical: that is because all the parties have, in effect, the same reasons and so they select the same principles" (*op. cit.*, p. 140, footnote 14).

Presumably, then, a genuine dialogical theory would be one featuring an account of the moral point of view according to which adopting it, and using it to identify the true principles of morality, are essentially collective acts. Insofar as the moral point of view is the standpoint from which the true principles of morality are revealed, it is essentially a collective standpoint. It must be entered with other people, and the true principles of morality will be revealed only after a certain kind of interaction in which everyone participates.

As we have seen, Habermas presents the contrast between Rawls's theory and his own in these terms. But we should examine in greater detail how the two theories are to be understood.

II

In some of his other works,[6] Habermas has proposed that any valid moral norm must satisfy the following condition:

> (U). *All* affected can accept the consequences and the side effects its *general* observance can be anticipated to have for the satisfaction of *everyone's* interests (and these consequences are preferred to those of known alternative possibilities for regulation) (*ibid.*, p. 65).

I take it that by a norm Habermas means a rule or principle, rather than an observable regularity in the behavior of people in certain kinds of situations.

It is natural to suppose that (U) is intended by Habermas as a formulation of the moral point of view, an alternative to Rawls's original position. (U) shares some features with what is perhaps the simplest account of the moral point of view, that associated with the monological device of universalization. One form of universalization directs the agent to consider whether she could accept everyone's acting in accordance with the particular principle (for a given kind of situation) that she proposes to adopt. (U) requires that not just the agent, but all affected, be able to accept everyone's complying with the principle for a given situation. Another form of universalization directs the agent to imagine herself on the receiving end of action in accordance with the principle she proposes to adopt, and to consider whether she could then accept that principle. If imagining oneself on the receiving end means taking on not only the circumstances of others, but also their desires and projects, this form of universalization generates a concern for the interests of all who will be affected by action on a given principle. (U) has a feature similar to this as well, in that it stipulates that each is to make a decision about whether a given principle should be generally adopted by considering how action on it would affect everyone's interests.

---

[6] "Discourse Ethics. Notes on a Program of Philosophical Justification" in *Moral Consciousness and Communicative Action*, Christian Lenhardt and Shierry Weber Nicholsen, trans. (Cambridge. MIT, 1990).

Let us explore more fully the similarities and differences between ordinary monological universalization and (U). Habermas supposes that each agent engages in some form of thinking that enables her to generate a conclusion regarding what would be in everyone's interests, but this does not distinguish his theory from one enjoining ordinary universalization. The distinctive feature of Habermas's view is that all those potentially affected must engage in such thinking, and the results they reach must agree. This raises the question of why they must agree, and I shall return to this. For present purposes, the important point is that Habermas realistically assumes that agreement among agents each of whom is attempting to determine what would be in everyone's interests will not be achieved without some process of interaction that brings what will undoubtedly be divergent initial conclusions into line. What does this process of interaction involve?

First, we should consider the thinking by which each reaches a conclusion concerning what would be in everyone's interests. Habermas speaks of each taking the perspective of everyone else, and this suggests that each may generate a view about what would be in everyone's interests by engaging in ordinary monological universalization. In monological universalization of the second sort described above, an agent takes the perspective of each person who will be affected, and arrives at an understanding of the principles that would be in everyone's interests by somehow combining the deliverances of all these perspectives. This involves the agent's *identification* with others—her making their concerns her own—and these acts of identification in turn *constitute* the impartiality characteristic of moral thinking: impartiality in the resolution of conflicts of interest. If we suppose that to universalize is to adopt the moral point of view, we can speak in this connection of a constitutive account of the moral point of view, one that posits a procedure the employment of which constitutes moral impartiality.

When the thinking by which we determine what would be in everyone's interests is understood in this way, in terms of monological universalization, the process of interaction that brings the parties into agreement will presumably be one in which they exchange information about the desires and projects of each, and about other relevant facts, until each has the same beliefs about these matters as the others. Given this common basis of belief, the results of their monological universalizations should agree. But it is important to note that the posited interaction will have a purely epistemic function. In particular, the fact that the interaction is a "joint" process plays no role in constituting the impartiality characteristic of morality. Moral impartiality is constituted by each agent's identification with all the others. The jointness of the enterprise merely eliminates a source of error that can corrupt this impartiality-constituting act.

This characterization of moral thinking and the role that interaction plays in it is not completely in accord with Habermas's view, however. He seems to envisage a form of interaction that involves not just the exchange of relevant information, but collective reasoning, a joint effort to assess the force of a body of relevant reasons. Presumably, these will be reasons germane to the question of what would be in everyone's interests, which will include reasons germane to the question of what the true interests of each are and reasons germane to the question of how conflicts of interest are to be resolved. The idea appears to be that as the process of collective reasoning unfolds, a clearer picture of what these reasons require will emerge. And as all become aware of this clearer picture, they will be brought into agreement about what would be in everyone's interests.

But again, the envisaged form of interaction will have only epistemic significance. On this second interpretation, moral impartiality is not constituted by an act of identification. It is constituted by certain reasons relevant to the resolution of conflicts of interest. We can follow Habermas in taking these to be reasons concerning what would be in everyone's interests. To give impartiality a place in one's thinking—to become someone who takes an impartial stance—is to grasp the force of these reasons. Thus, the envisaged interaction goes beyond the elimination of factual ignorance to encompass a collective effort to assess the rational significance of reasons of a certain sort. But the goal of the process is nevertheless to overcome various cognitive shortcomings that have the potential to corrupt our appreciation of the import of the conceptual resource, a certain body of reasons, that provides the true ground of impartiality. The interaction does not itself constitute this impartiality.

We can summarize these points by distinguishing two different ways of understanding the moral point of view, the standpoint from which the true principles of morality are revealed. As I have noted, the moral point of view can be interpreted *constitutively*, as a standpoint the adoption of which constitutes moral impartiality. Ordinary monological universalization of the second sort, in which an agent adopts the perspectives of all affected by identifying with them and their concerns, is an example. Alternatively, the moral point of view can be interpreted *epistemically*, as a standpoint the adoption of which enables us to reduce or eliminate various distortions that can affect our grasp of whatever conceptual resources ground moral impartiality.

On an epistemic interpretation, as well as on a constitutive one, adopting the moral point of view can be understood as involving ideal role taking. One way to eliminate, or at least reduce, the characteristic distortions affecting our grasp of moral impartiality is to put ourselves in the positions of those who will be affected, so as to see how the impartiality constituting conceptual resources look from their points of view. We could even say that in doing this, we become impartial in a further respect: we eliminate bias in our assessment of the import of the relevant conceptual resources. But this way of becoming more impartial does not by itself constitute the impartiality characteristic of moral thinking, impartiality in the resolution of conflicts of interest. That form of impartiality is constituted by a certain thought experiment, or by (the grasp of) reasons relating to what would be in everyone's interests.[7]

An epistemic account of the moral point of view could be employed in conjunction with a constitutive account. We would then have a double employment of ideal role taking. Each party would identify with those affected in order to constitute himself as impartial, and then employ the epistemic tactic of adopting the perspectives of other people performing these acts of identification in order to overcome the error about the true principles of morality that disagreement reveals to be present somewhere in the group.

As we have seen, however, it does not appear that Habermas's theory envisages the constitution of moral impartiality by acts of identification. Rather, moral impartiality is constituted by reasons governing the resolution of conflicts of interest, reasons relating to what would be in everyone's interests. Ideal role taking—the process of entering into the perspectives of others—then involves only an effort to overcome bias in the assessment of the force of these reasons. This could be a monological process, in which each individual imagines himself in the positions of the others who will be affected and considers how the relevant reasons look from those positions. But Habermas's emphasis on dialogue is readily explained. The exchange of arguments concerning the force of the relevant reasons reveals and corrects additional bias. Indeed, the exchange of arguments itself can be regarded as involving the adoption by each party of the perspectives of the others. To offer an argument is to offer one's perspective on the reasons, and in coming to understand this argument, others, in effect, adopt one's perspective. Still, this process has an epistemic, rather than constitutive, significance. Moral impartiality is constituted by (the grasp of) the relevant reasons, and adopting the perspectives of the others—either by a monological act of imagination or by the exchange of arguments—is just a way of overcoming distortions that can affect our understanding of the import of these reasons.

---

[7] I discuss these two ways of understanding the moral point of view in more detail in *Collective Rationality and Collective Reasoning* (New York: Cambridge, 2001), chapter 6.

An epistemic interpretation of the moral point of view can, then, take either a monological or a dialogical form. Since, as will be discussed in the next section, there are reasons why agreement on the true principles of morality would be desirable, there may be a presumption in favor of the dialogical form. It is more likely to produce agreement. But the dialogical form retains the characteristic features of an epistemic interpretation of the moral point of view. It counts as a standpoint from which the true principles of morality are revealed only in the sense of being a standpoint from which error about what is required by the reasons relevant to the identification of these principles can be overcome. Moral impartiality is constituted by the reasons.

III

It will be useful to explore some further aspects of Habermas's theory in light of these observations.

(U) establishes being the object of an agreement of a certain kind as a condition that any genuine moral principle must satisfy. But the agreement must apparently comprise all moral agents—since all are potentially affected by any agent's adoption of a principle—and it seems that there are few, if any, principles that are accepted by all moral agents. So if (U) is interpreted as requiring an actual agreement at the present time, the result will be that few, if any, of the principles on which we might act can be viewed as genuine principles of morality.[8]

---

[8] For discussion of this point, see Michael Rosen, "Utopia in Frankfurt: Realism and Optimism in the Life's Work of Jurgen Habermas" (a review essay discussing two recent works by Habermas), *The Times Literary Supplement*, n 5036 (October 8, 1999), pp. 3-4.

There may, however, be a way of avoiding this difficulty. (U) speaks of what all affected *can* accept, and this might be glossed as what they *could* accept under certain counterfactual conditions. Habermas[9] provides a definition of the term 'discourse' which introduces such conditions. He says that he will speak of discourse when the meaning of the claim at issue forces the participants to suppose that a rationally motivated agreement could be achieved if the argument were conducted openly enough and continued long enough (*ibid.*, p. 42).[10] If we can avail ourselves of this device, we can regard (U) as stating a condition that could be satisfied by some of our practical principles now. They may be the very principles on which all moral agents would agree if argument were conducted openly enough and continued long enough.

It should be noted that (U) does not specify that the satisfaction of the condition it formulates must be achieved by argument. The condition articulated by (U) is a condition that a principle satisfies by virtue of being the object of agreement among all affected, but (U) is silent on the method that is to be employed in attaining this agreement. In theory, it could be attained by a process of bargaining in which each makes concessions from his most preferred principle or set of principles until a compromise acceptable to all is found. To accept this compromise would be to accept an interpretation of what would be in everyone's interests. Habermas is clear, however, that he envisages a process of collective *reasoning* rather than *bargaining*.[11] One reason for this is that he wants to ensure that our understanding of the true principles of morality is not corrupted by various forms of social power. Agreement is attained when collective reasoning overcomes all the distortions that may affect how the members of the designated group perceive the import of a body of reasons (concerning what would be in everyone's interests) that should have the same significance for each of them. This gives us a way of interpreting Habermas's talk of the emergence of an ideally extended we-perspective out of everyone's efforts to adopt the perspectives of the others. The we-perspective is implicit in the process of collective reasoning, as the condition that will obtain after the exchange of arguments eliminates distorted perceptions of the import of the relevant reasons, thus giving all the same understanding of the force of these reasons.

---

[9] *The Theory of Communicative Action*, Volume 1, Thomas McCarthy, trans. (Boston: Beacon, 1984).

[10] See also "Discourse Ethics," p. 105.

[11] Habermas distinguishes the communicative action associated with moral discourse from strategic action, which involves attempts to influence the other parties by offering rewards or threatening punishments, see "Discourse Ethics," p. 58.

Another issue that deserves attention concerns why the parties regard what would be in everyone's interests as providing the reason on the basis of which norms or principles are to be accepted. We could say that this is because they are reasoning morally—that is, about how conflicts of interest are to be resolved—and in such cases (by definition, as it were) the relevant reasons relate to what would be in everyone's interests. But Habermas appears to envisage a deeper explanation. To see this, we must first consider a prior question. Why do the parties need to reach agreement? The explanation that comes most naturally to mind involves the role of agreement in securing certain cooperative benefits. Agreement on principles governing social interaction, especially in situations where interests conflict, is required for harmonious living generally.[12] In light of this, the answer to the question of why the parties must reason in terms of what would be in everyone's interests is straightforward. Agreement can be attained by a process of reasoning only if each party presents reasons of a kind that the others could accept, and no party can suppose that the others will be willing to cooperate on the basis that her interests are more important than theirs. So the imperative to find a framework for voluntary cooperation in the face of conflicting interests forces the parties to reason in terms of what would be in everyone's interests.[13]

This last point might seem to allow us to offer a kind of constitutive interpretation of Habermas's theory after all. Moral impartiality is constituted by the need to find a basis of cooperation acceptable to all. But this would be inaccurate. We might be able to take such a constitutive view of the process by which moral principles are identified if the relevant agreement were understood as a compromise reached by bargaining.[14] If agreement is to be attained by a process of reasoning, however, we cannot regard the need to find a common basis of cooperation as constituting moral impartiality. This need merely forces each to look for reasons acceptable to all. It does not ground the existence of these reasons.

---

[12] For Habermas, communicative action is "action oriented to reaching an understanding." The understanding he has in mind is specifically an understanding about what to do that can provide a basis for social coordination; see *The Theory of Communicative Action*, pp. 94-101.

[13] For an interpretation of Habermas that stresses this point, see Jon Elster, "The Market and the Forum: Three Varieties of Political Theory," in Elster and Aanund Hylland, eds., *The Foundations of Social Choice Theory* (New York: Cambridge, 1986), pp. 103-32; reprinted in James Bohman and Rehg, eds., *Deliberative Democracy: Essays on Reason and Politics* (Cambridge: MIT, 1997), pp. 3-33.

[14] For a view of this sort, see Gilbert Harman, "Moral Relativism Defended," *Philosophical Review*, LXXXIV (January 1975): 3-22. Harman understands moral judgments to be relative to motivational attitudes that take the form of intentions to adhere to an agreement on the understanding that others (in the relevant com-

The thinking here should, perhaps, be explained more fully. It is plausible that reasoning can, to a certain extent, shape—that is, alter—the content of the reasons with which it works. But if the process by which moral principles are identified is not to be corrupted by disparities in social power, any shaping of reasons that is involved must be understood as itself displaying a sensitivity to something normative that guides everyone engaged in it.[15] If it is understood solely as process of construction, there will be nothing to prevent disparities in social power from being manifested in the result.

These additional elements of Habermas's theory do not, then, force us to revise the conclusions reached above about the essentially epistemic nature of the interaction he envisages. Moral impartiality is constituted by the relevant reasons, those relating to what would be in everyone's interests. The deliberative process—the process of exchanging arguments—can be viewed as one by which the perspective of each is made known to the group as a whole, thus providing a way that each can enter into the perspectives of the others. But this must be understood simply as an epistemic procedure that facilitates the attainment by all of an accurate grasp of the relevant (impartiality-grounding) reasons. It is the potential of this process of entering into the perspectives of others to yield a common understanding of what the relevant reasons require that enables us to speak of a we-perspective in connection with it.

As was noted earlier, however, collective reasoning can also provide epistemic benefits in the context of Rawls's method of reflective equilibrium. This is a method for tracking the force of certain reasons, reasons that guide us in the construction of a monological

---

munity) intend to adhere to it as well. The precise form of the agreement is determined by bargaining, in which people bring to bear different amounts of bargaining power. So the interpretation of impartiality in the resolution of conflicts of interest that is constituted by the agreement reflects this fact.

[15] Elsewhere, I explore the possibility that reasoning can shape the reasons it works with—*Collective Rationality and Collective Reasoning*, chapter 5, section V, and chapter 7, section V. As I note in that discussion, this possibility is connected with the idea that normativity can have a history—that what constitutes correct thought and action can change as history unfolds, and thus be different now from what it was in the past.

account of the moral point of view: that associated with the original position. The original position provides what I have called a constitutive interpretation of the moral point of view. The veil of ignorance forces one to identify with everybody. It is plausible to suppose that when those employing the method of reflective equilibrium offer their reasoning for criticism by others, especially under the idealized conditions associated with the Habermasian notion of discourse, the tracking of the relevant reasons will be more accurate, and at the limit, result in agreement (about the proper characterization of the original position and what would be chosen there). The only real difference between the theories concerns how the conceptual resources that constitute impartiality are understood. Habermas understands them as an open-ended body of reasons relating to what would be in everyone's interests, while Rawls offers an explication of moral impartiality in terms of a certain constitutive account of the moral point of view.

IV

It might be objected that we must understand Habermasian dialogicality as having a further dimension if the outcome of collective reasoning is to satisfy the condition stated by (U). In Habermas's view, the process of seeking, through collective reasoning, an understanding of what is required by the reasons relevant to a given question commits us to accepting certain constraints on the process of inquiry—basically, constraints designed to ensure that the best argument prevails, such as a requirement that everyone be given a chance to speak and a respectful hearing. Habermas supports this claim by arguing that the pragmatic presuppositions of discourse enjoin us to obtain the views of all who might have relevant evidence.[16] And when moral questions are being addressed, the opinions of others concerning what would be in everyone's interests are always potentially relevant evidence. This, in turn, is supposed to justify the general acceptance of (U) as a condition that must be satisfied by moral principles.

One possible interpretation of the connection between the specified constraints on deliberative interaction and satisfaction of the condition stated by (U) is that the constraints make the correct appreciation of the relevant reasons more likely, thus—in the ideal case—producing agreement. That is, (U) describes the situation that will obtain after the epistemic benefits of collective reasoning concerning what would be in everyone's interests, carried out under conditions designed to maximize its epistemic potential, have been fully realized. It provides a general characterization of the outcome that will be attained when a group succeeds in constructing a we-perspective on the relevant reasons by a process that involves entering, through argument, into the perspectives of others.

---

[16] "Under the pragmatic presuppositions of an inclusive and noncoercive rational discourse among free and equal participants, everyone is required to take the perspective of everyone else, and thus project herself into the understandings of self and world of all others"—"Reconciliation," p. 117. Presumably this projection is to be accomplished by comprehending their arguments.

Habermas's understanding of the relation between the constraints on deliberative interaction associated with the notion of discourse and acceptance of (U) appears, however, to have a further aspect. In one place,[17] he speaks of (U) as a *rule of moral argument* that is implied by the constraints. When what is at issue is the elimination of cognitive malfunction, we can regard agreement as an endpoint implicit in the process of collective reasoning. But it would be odd to describe the attainment of agreement as enjoined by a rule. Agreement would just reflect a convergence of judgments made by people independently trying to determine what the relevant reasons support. To speak of agreement as a rule, we must suppose that there is some further constraint requiring that the judgments that conclude the relevant reasoning be explicitly collective acts. That is, we must suppose that each is constrained to suspend judgment until a conclusion all can accept has been found.

This, of course, raises the question of what could ground such a constraint. One suggestion, supported by some things Habermas says, is that the participants in discourse have an obligation to justify any judgment they make concerning what ought to be done *to* those who will be affected. But we must be clear about how this claim is to be understood. If it merely means giving reasons for one's ought-judgments that *should* convince others (if they are competent reasoners), no requirement to suspend judgment as long as anyone disagrees will be implied. To obtain such a requirement, the stipulation that one must justify one's ought-judgments to those who will be affected must be understood in such a way that their actual assent is required. But what justification could there be for adopting this posture? If we say that the obligation to justify our judgments to others is a free-standing moral requirement, independent of and prior to (U)—a requirement to interact with others only on a basis they can accept—then there will be at least one moral principle the validity of which is not grounded in (U).[18]

---

[17] See "Remarks on Discourse Ethics," in *Justification and Application*, Ciaran Cronin, trans. (Cambridge: MIT, 1993), pp. 19-111, here p. 32.

A different answer is that obtaining the assent of others is not a *moral* but a *cognitive* requirement. When we are seeking to determine what is in the interests of all, the relevant considerations are provided by the needs of each individual, and each person has a kind of privileged access to what counts as a genuine need of hers.[19] Thus one must defer to the judgments of others concerning what their needs are, at least in the sense of suspending judgment as long as they dissent. The result is that only a collective judgment to which each makes a contribution can capture the true force of the *full* set of relevant (need-based) reasons. If this way of viewing the matter is correct, the requirement that all accept the conclusion reached will follow from the nature of the reasons relevant to the question being considered (what would be in everyone's interests), irrespective of whether agreement on the answer is necessary for voluntary social cooperation. It can thus plausibly be construed as a rule of argument for questions of that kind.

This proposal has the potential to fill what may be a gap in Habermas's theory. It is often said that Habermas's invocation of the pragmatic presuppositions of discourse yields a kind of transcendental deduction of morality, showing it to be a condition of the possibility of our lives as reasoning animals. As Habermas puts it in his exchange with Rawls: "the moral point of view is already implicit in the socio-ontological constitution of the public practice of argumentation, comprising the complex of relations of mutual recognition that participants in rational discourse 'must' accept (in the sense of weak transcendental necessity)" (*op. cit.*, p. 127).

But while giving everyone a chance to speak and a respectful hearing may be a condition of the possibility of rational inquiry within a group, this yields only an "internal" morality of the process of reasoning itself, a set of requirements governing how the interlocutors must treat each other if the cognitive benefits of discourse are to be achieved. What appears to be required for the posited deduction of morality from the presuppositions of discourse is a demonstration that a moral relationship in the *rest* of our lives is a condition of the possibility of rational discourse. But if, when moral issues are at stake, the force of the relevant reasons must be captured in a collective judgment, so that each must suspend judgment until a conclusion all can accept has been found, we can obtain the desired result. Entering into a relationship that could be regarded as having a moral character even if achieved by nondiscursive means, a kind of solidarity, will be a necessary part of proceeding rationally when moral issues are being confronted—when decisions are to be made concerning the proper resolution of conflicts of interest.

---

[18] See Habermas's discussion of the *Bindungseffekt* ("Discourse Ethics," pp. 58-59), which is based on the idea that a speaker has an obligation to redeem discursively the claim the hearer is to accept. I have suggested that it is problematic to regard these obligations as moral. But if they rather derive from the rules of the language game of making ought judgments (which require one to present the reasons supporting such a judgment when asked), there seems to be no basis for supposing that the obligation is only fulfilled if others actually accept the justification.

[19] Rehg appears to make a proposal of this sort in "Discourse and the Moral Point of View: Deriving a Dialogical Principle of Universalization," *Inquiry*, xxiv (1991): 27-48. See also his *Insight and Solidarity: The Discourse Ethics of Jürgen Habermas* (Berkeley: California UP, 1994), chapter 3.

I have said that this approach *may* fill a gap in Habermas's theory because it is not entirely clear that he envisages such an extension to the rest of our lives of the moral import of the presuppositions of rational discourse. Habermas observes that Rawls has criticized his transcendental argument on the ground that it could not yield a sufficiently structured theory of justice. Habermas's reply is that he subscribes to a division of labor between moral theory and the theory of action, and that "the conceptual structuring of the contexts of interaction to which questions of political justice refer is not within the province of moral theory"; he then adds that, "together with the content of action conflicts in need of resolution, a whole conceptual frame for normatively regulated interaction is forced upon us" (*op. cit.*, p. 127).

One interpretation of what Habermas is saying here is that the presuppositions of discourse bring with them only an internal morality of discourse, and further conceptual structures, which inevitably present themselves when we encounter conflicts of interest, constitute the true ground of moral impartiality. In this case, he would be largely agreeing with the interpretation I have been offering. When he says that the moral point of view is implicit in the public practice of argumentation, he would have in mind only an epistemic construal of this point of view, as a point of view the adoption of which eliminates bias in the assessment of the import of independent conceptual structures. Moral impartiality per se would be constituted by these independent structures.

If this is Habermas's position, however, his way of presenting it is at least misleading. He is no doubt correct that the constraints on discourse that are necessary if it is to achieve its cognitive goal of better justified judgment require a certain respect for others and thus have a moral character. But if his view is that moral impartiality per se is constituted by independent structures, the best interpretation of the morality of discourse is that the constraints that must be respected if discourse is to achieve its cognitive goal resemble in certain respects the paradigmatically moral constraints associated with the independent structures. The constraints on discourse do not serve as the point at which the notion of moral impartiality—impartiality in the resolution of conflicts of interest—enters our thinking.

To return to the suggestion that a collective judgment is required to capture the reasons relevant to the identification of the true principles of morality, we should be clear about how it fits into the framework I have been sketching. It does not force an abandonment of the central claim of the epistemic construal of the role of collective reasoning in our moral life: that the true principles of morality are constituted as such by their relation to reasons that are relevant to their adoption, reasons that constitute the ground of moral impartiality, and collective reasoning is simply an effective way to eliminate various impediments to our attempts to track the force of these reasons. The distinctive feature of the proposal concerning collective judgment is that given the assumption it makes about the epistemic status of the relevant (need-based) reasons—namely, that each has a kind of privileged access to some of them—an individual reasoning alone cannot make adequate epistemic contact with all of these reasons. She can correctly assess the force of the full set of relevant reasons only if she is prepared to suspend judgment when others assert that general acceptance of a given moral principle would frustrate certain of their needs. Thus, a judgment regarding what is required by all the relevant reasons must be collectively generated.

Still, this is a stronger claim than can be made in connection with Rawls's theory. While those employing Rawls's method of reflective equilibrium would normally benefit from reasoning collectively, it is possible that an individual reasoning independently could gain adequate epistemic access to all the relevant reasons. So by assuming that each has a kind of privileged access to her own needs, at least in the sense that others must suspend judgment about what they are as long as she disagrees, we can view Habermas's theory as importantly different from Rawls's, despite the fact that moral impartiality is constituted by the relevant reasons, and the envisaged interaction is merely a way of capturing their true force.

The assumption of a kind of privileged access is, however, highly questionable. It is plausible that each of us can, in principle, grasp any concept that can be grasped by anyone else. Philip Pettit[20] has characterized this feature of concepts and rules as *commonability*. As he puts it: "Commonability is a property of rules. ... [I]t requires them to be such that if there are others, and if they try to identify the rules [one] follows as rules they can follow themselves, then it is possible they will succeed: it is possible that they will knowledgeably identify the rules in that way" (*ibid.*, p. 181). If Pettit is right that our concepts and rules are commonable, there are no reasons to which individuals have privileged access, even in the weak sense that others must suspend judgment as long as they disagree. But then collective reasoning about what would be in the interests of all will have no special epistemic features distinguishing it from collective reasoning directed toward other questions. It will be a matter of the joint elimination of cognitive malfunction in the assessment of the force of a body of reasons all of which are in principle accessible to everyone. And agreement will be achieved when the last independent judgment about what these reasons require comes into line with the rest.[21]

---

[20] *The Common Mind: An Essay on Psychology, Society, and Politics* (New York: Oxford, 1993).

V

It may be useful, as a way of summarizing this discussion of Habermas's theory, to return to the suggestion that (U) can be understood as an account of the moral point of view. (U) stipulates that there must be (a) agreement, regarding (b) what would be in everyone's interests. It is natural to suppose that *both* these elements enter into (U)'s claim to articulate the impartiality and equal consideration essential to morality. But while there is an obvious connection between the notion of what would be in everyone's interests and the basic moral concepts of impartiality and equal consideration, the status of the stipulation that there be agreement about what would satisfy this condition is problematic.

One way that agreement about what would be in everyone's interests could be achieved is by a process of bargaining and compromise, in which each makes concessions until something all are prepared to accept has been found. But as has been noted, Habermas rejects the suggestion that agreement about the true principles of morality is to be reached in this way. It is rather to be reached by a process of reasoning, where the relevant reasons are precisely those germane to the determination of what would be in everyone's interests. If this is how the agreement stipulated by (U) is to be understood, however, the connection this principle has with impartiality and equal consideration is provided solely by (b)—that is, by the concern with the interests of all that each person is assumed to have (by virtue of attempting to track the force of the relevant reasons). The stipulated agreement is of purely epistemic significance. It is simply a sign that the group has attained a correct understanding of what would be in everyone's interests.

---

[21] The result will thus be what I call a *piecemeal consensus*. I explain this notion in "Discourse and Morality," *Ethics*, cx (April 2000): 520-21, and in *Collective Rationality and Collective Reasoning*, p. 112.

In "Discourse and Morality," I present an argument to the effect that the Wittgensteinian considerations usually thought to show that rule-following cannot be too atomic also tell against the possibility of a collective judgment involving all moral agents. This strong result is not, however, necessary to establish that there is no cognitive requirement to suspend judgment as long as anyone disagrees. We can concede that there can be legitimate collective judgments involving all moral agents, provided that in any such case, the corresponding judgment made by an independent individual could also be rationally in order. In *Collective Rationality and Collective Reasoning*, chapter 5, I propose a hybrid account of reasons—which combines practices with an independent source of normativity—that works this way.

Given these observations, Habermas's account of the moral point of view can actually be combined with Rawls's. The idea of being in everyone's interests requires an interpretation, which Habermas does not provide. He may feel that this will be worked out as shared deliberation proceeds. But the agreement about what would be in everyone's interests which is necessary (if the benefits of voluntary social cooperation are to be realized) is most likely to be achieved if the parties interpret this concept in terms of a fair resolution of their competing interests—if they treat what would be in everyone's interests as what would be fair. We can thus understand the process leading to the satisfaction of (U) as a process of collective reasoning in which a group of moral agents seeks to track the force of reasons of fairness. If agreement results, this could be taken as a sign that they have attained a correct appreciation of what fairness requires. Earlier, it was mentioned that the notion of being in everyone's interests could be understood in terms of a compromise achieved by a bargaining process. But this would reflect the influence of various forms of social power. If, by contrast, the parties deliberate about what reasons of fairness require, the process in which they engage to identify the true principles of morality will be guided by something that has the same force for all. Thus they will be able to converge on a way of reconciling their competing interests which is not corrupted by disparities in power.

But if we understand what would be in everyone's interest in terms of fairness, there will be little difference between Habermas's theory and Rawls's. Rawls's theory of justice as fairness purports to show that a certain account of the principles of justice is supported by reasons of fairness. And as we have seen, the method associated with this theory—the method of reflective equilibrium—can be employed in a way that gives collective reasoning an epistemic role. The only difference between Habermas and Rawls will be that Rawls has taken up the challenge of explicating, with the help of his device of choice in the original position, the notion of a fair resolution of the competing interests. He has employed a monological, constitutive account of the moral point of view to explicate the reasons of fairness that it is plausible to suppose Habermas's dialogical—but epistemic—interpretation of mutual perspective taking must treat as the focus of moral deliberation.[22]

It does not follow that Habermas should be criticized for declining to provide something similar. It appears that one of the reasons he is dissatisfied with Rawls's theory is that he thinks that by using the original position to explicate moral impartiality, Rawls has closed off various routes by which considerations germane to the determination of what would be in everyone's interests might be received.[23] He prefers to think that we need to remain open, in the course of ongoing deliberation, to the possibility of learning something new about what would be in everyone's interests.

But this is not a serious disagreement. Habermas is right that deliberation must continue, but this will involve a collective effort to come to a better understanding of the import of the conceptual resources that ground moral impartiality. And attempts to explicate moral impartiality with a constitutive account of the moral point of view are certainly a legitimate part of the ongoing process of inquiry. In light of our proclivity for generalizing about matters of interest to us, proposals of this sort are to be expected. Thus, inquiry into what would truly be in everyone's interests will normally take the form of a dialectical alternation between attempts to capture in general terms the essence of moral impartiality—with constitutive accounts of the moral point of view or in other ways—and the criticism of these attempts as inadequate. We should not regard this alternation as a rut to break out of. Our understanding of what morality requires advances as a result of both parts of the process.

I conclude, then, that the idea that there is an important distinction to be made between the moral theories of Rawls and Habermas—and more broadly between monological and dialogical moral theories—comes to naught. In one sense, all theories that make a place for moral reasoning are dialogical. All can take advantage of the epistemic benefits of collective reasoning in assessing the force of the reasons they regard as grounding moral impartiality. But there does not appear to be any way of viewing moral impartiality per se as constituted by an essentially interpersonal mechanism unless it is a mechanism of bargaining. If reasoning is rather involved, moral impartiality is grounded in the relevant reasons, and the parties become impartial, whether individually or collectively, by grasping the force of these reasons.

---

[22] Rawls says, "it is you and I—and so all citizens over time, one by one and in associations here and there—who judge the merits of the original position as a device of representation and the principles it yields"—"Reply to Habermas," p. 140, footnote 14. And also, "[reflective] equilibrium is fully intersubjective: that is, each citizen has taken into account the reasoning and arguments of every other citizen (p. 141, footnote 16.)

[23] See "Reconciliation," part I.

This has the consequence, however, that the construction of a we-perspective must be understood epistemically. Each individual adopts the perspectives of all others by coming to understand their arguments. When this process is carried out under ideal conditions, the result is common understanding of what the relevant, impartiality-grounding, reasons support—and thus of what the true principles of morality (of justice) are.

CHRISTOPHER MCMAHON

University of California/Santa Barbara

DE FACTO DEPENDENCE[*]

Two friends throw rocks at a window; Suzy's rock hits the window, while Billy's sails harmlessly through the now empty frame. It was Suzy's throw, not Billy's, that caused the window to break. But counterfactually, the two throws seem on a par. Had neither occurred, the window would not have broken; had either occurred without the other, it (still) would have. This is an instance of the *preemption problem* for dependence-style accounts of causation. Suzy's throw is a cause despite being no more depended on than Billy's.

But is it really so that Suzy's throw is no more depended on than Billy's? Take away Billy's throw while holding other things fixed, and the window still breaks. But if we take away Suzy's throw while holding other things fixed, it does not break. This is because one of the things we have to hold fixed is that Billy's rock never touches the window. Had Suzy not thrown, with Billy's rock never making contact, the window would have had nothing to fear from either rock, so it would not have broken.

Effects do not always counterfactually depend on their causes. But they do (seem (so far) to) *de facto* depend on them—to depend on them with the right things held fixed. Of course, a story will be needed about what the "right things" are. That is the business of the next few sections, after which I compare the de facto theory to some other recent responses to preemption.

I ANALOGY

Preemption is a problem for the counterfactual theory of causation, familiar from the work of David Lewis.[1] Compare the problem facing another counterfactual theory, also discussed by Lewis. One starts out thinking that

> x is disposed at time t to give response r to stimulus s if and only if, if x were to undergo stimulus s at time t, x would give response r.[2]

---

[*] I am grateful to David Christensen, Cian Dorr, Christopher Hitchcock, Igal Kvart, Krista Lawlor, David Lewis, Laurie Paul, Judea Pearl, and Ana Carolina Sartorio. Ned Hall helped the most. Thanks also to audiences at University of Toronto, University of Vermont, UCLA, Stanford University, and Princeton University The theory presented here is probably closest to Hitchcock, "The Intransitivity of Causation Revealed in Equations and Graphs," this JOURNAL, XCVIII, 6 (June 2001): 273-99; and to Pearl, *Causality* (New York: Cambridge, 2000), chapter 10.

[1] "Causation," this JOURNAL, LXX, 17 (October 11, 1973): 556-67, reprinted with "Postscripts" in *Philosophical Papers, Volume 2* (New York: Oxford, 1986), pp. 241-69. See also Ardon Lyon, "Causality," *British Journal for the Philosophy of Science*, XVIII (1967): 1-20

[2] Lewis, "Finkish Dispositions," *Philosophical Quarterly*, XLVII (1997): 143-58. See also Charles B. Martin, "Dispositions and Conditionals," *Philosophical Quarterly*, XLIV

0022-362X/02/9903/130-48

But, Lewis notes, "stimulus s...might chance to be the very thing that would cause the disposition...to go away. If it went away quickly enough, it would not be manifested."[3] This is the *finkishness problem*, so called because the theory overlooks dispositions that "fink out" when the stimulus is applied.[4] An example is Saul Kripke's[5] killer yellow, which destroys the brains of those who set eyes on it before any visual experiences can occur.

Finkishness and preemption have a great deal in common.[6] Both make trouble for analyses of the same basic form: x is F if and only if $(Ax > Bx)$. Both arise because x is F in virtue of some underlying G, yet x would not have been G, had it been A. Where the problems differ is in the identity of F and G. Finkishness has $F =$ x's disposition and $G =$ its grounds; preemption has $F =$ x's causing of y and $G =$ ...well, that is what we have to explain.

Socrates distinguishes between "the cause" of an effect, and "that without which the cause would not be a cause" (*Phaedo* 98e). Take the fact that Billy's rock never touches the window. This is not a cause of the window's shattering; it does not contribute to the shattering in any way. But it does help to make something else the cause. It is a cause *maker* (read like "king maker").[7] The no-show status of Billy's rock helps to make Suzy's throw the cause by putting the effect in need of it. So, proposal: where finkishness has

$F =$ x's dispositional property
$G =$ the disposition's grounds, that without which s would not be sufficient for r

preemption has

$F =$ x's causal relation to y
$G =$ the relation's grounds, that without which x would not be required by y

Preemption happens because to take away the cause is, sometimes, to take away more. It is to take away one of the reasons it *is* a cause,

---

(1994): 1-8; and Robert K. Shope, "The Conditional Fallacy in Contemporary Philosophy," this JOURNAL, LXXV, 8 (August 1978). 397-413.

[3] "Finkish Dispositions," p. 144.

[4] There is a parallel problem, Lewis notes, about finkish undispositions: previously undisposed x might acquire from s the disposition to respond to s with r.

[5] This example is from unpublished lectures on secondary qualities.

[6] They are usefully juxtaposed in John Collins, "Preemptive Preemption," this JOURNAL, XCVII, 4 (April 2000): 223-34.

[7] Cause makers are called "ennoblers" in my "Advertisement for a Sketch of an Outline of a Proto-Theory of Causation," to appear in Collins, Ned Hall, and Laurie Paul, eds. *Causation and Conditionals* (Cambridge: MIT, forthcoming).

namely, that other would-be causes miss the mark, leaving *y* with no other source for what *x* provides. The effect cannot be expected to follow *x* out of existence, if the reasons for its dependence on *x* go out of existence first!

The claim so far is that preemption is *analogous* to finkishness. One might try to turn the analogy into an identity. What is there to *x*'s causing of *y*, beyond the fact that *x* and *y* both occur? A *disposition* on the part of the circumstances to lose *y* if "stimulated" by the loss of *x*.[8] For *x* to cause *y* is for it to engage with that disposition. Whatever grounds the disposition contributes thereby to *x*'s status as cause.

II. HOLDING FIXED

No one is happy with existing theories of preemption. But about finkishness there is a certain amount of optimism. The form of the solution is thought to be this: *x* is disposed to produce *r* in response to stimulus *s* if and only if

> Were *x* to be presented with *s* while retaining grounding property *G*, then *x* by virtue of being *G* would give response *r*.[9]

Now, though, one must specify *G*. One does not want to do it in terms that presuppose the notion of a disposition, although if the presupposing was minor and indirect, we might still be interested. Best would be a definition that did not appeal to dispositionhood at all. Let us review some of the options.

(1) *G* = *x's total intrinsic nature.* This holds too much fixed. Some dispositions cannot be manifested without their bearers changing intrinsically. (Lewis gives the example of elasticity.) It also holds too little fixed, for some dispositions are grounded in extrinsic properties. An example is weight as conceived by operationalists: the disposition to register thus and so many pounds when put on a scale. This must mean on a *local* scale, since things weigh less on the moon than on Earth. But perhaps my locality would change if someone tried to weigh me. (You would have to chase me to the moon to get me on to that scale.) The disposition is there, but because it is grounded in external circumstances that change when the stimulus is applied, the counterfactual fails.[10]

(2) *G* = *x's total categorical nature.* This again holds too much fixed. Some dispositions are manifested through categorical changes in

---

[8] Why not also a disposition to "gain" *y* if stimulated by the addition of *x*? This idea has some merit, I think, but let us stick for now with the negative disposition.

[9] Lewis, "Finkish Dispositions", Collins, "Preemptive Preemption."

[10] Extrinsic dispositions are further discussed in my "Intrinsicness," *Philosophical Topics*, xxvi (1999): 479-504; see especially pp. 493-95.

their bearer: changes of shape and size, for instance, in the case of elasticity. The proposal also holds too little fixed, since some dispositions are grounded in hypothetical properties. Glass is fragile because of the way its chemical bonds are disposed to break under pressure. Warfarin is poisonous because of its anticoagulant powers.

(3) G = *any property of* x *such that were* x *to undergo* s *while retaining* G, *it would give response* r. No, because any x that is not undergoing s has a property like that: the material conditional property of producing r if undergoing s. A match is not disposed to bark like a dog on being struck. But that is what it would do, if we held fixed its property of either barking like a dog or not being struck.

(4) G = *any natural enough property of* x *such that were* x *to undergo* s *while retaining* G, *it would give response* r. The problem is not just that 'natural enough' is so obscure. It is that we get false positives even with quite natural Gs. Suppose this soft-boiled egg had been dropped, with its actual ovoid shape held fixed. The energy of the fall would have gone into the floor; the linoleum would have sagged and sprung back, flinging the egg into the air. The egg would have bounced even though it was disposed to break.

Option (4) yields not enough clear verdicts, and too many clearly wrong verdicts. Should we conclude that naturalness is no help? That would be too quick. Suppose that the Gs eligible to be held fixed could be identified (never mind how) using *comparative* naturalness rather than *absolute*. This would help with the first problem because whether G is more natural than H is a clearer matter than whether G is simply natural. It would help with the second problem because even a quite natural property can be *less* natural than some competitor. Our problems would be solved, then, if Gs arguing in favor of a dispositional link between s and r could be played off against Hs arguing against such a link, with the more natural condition prevailing. I do not know how to do this for finkish dispositions in general, but when it comes to the special case of preemption, a strategy does suggest itself.[11]

---

[11] I propose to postpone (= ignore) the question of what is the best thing to mean by 'natural'. This is partly because I am uncertain about it, in particular about the extent to which cognitive and cultural factors are allowed to come in. I get the term 'natural' and the idea that naturalness has a place in the metaphysician's toolkit, from Lewis—see his "New Work for a Theory of Universals," *Australasian Journal of Philosophy*, LXI (1983): 343-77, and (with Rae Langton) "Defining 'Intrinsicness'," *Philosophy and Phenomenological Research*, LVIII (1998): 333-45.

III MAKE OR BREAK

Causation involves a disposition on the part of the circumstances to respond with the loss of e to a stimulus consisting in the loss of c. One identifies circumstances as so disposed by playing cause makers G off against Hs arguing for an opposite verdict: cause *breakers*, let us call them.[12] G is a cause maker if and only if (a) it actually obtains (G is a fact), and (b) had it been that $\neg Oc$ & G, it would have been that $\neg Oe$ (for short, $\neg Oc >_G \neg Oe$[13]). One can see how it would create a presumption in favor of c's causing e that e depends on it modulo a factual condition G. What H might do to undermine the presumption, however, is quite unclear. I propose to sneak up on the answer slowly.

Normally in thinking about causation, we conceive of c as supplying something that e has need of, and would not otherwise get. Preemption makes matters a little more complicated, but only a little. Other events send the needed something out along alternative routes. Holding G fixed blocks those routes, so we are returned to the previous case: c makes an essential contribution, without which e would not have occurred.

But although that is how it is supposed to work, the requirement of dependence modulo G supports an almost opposite scenario. The effect was going to happen anyway, when c came along to threaten it. It did this by creating a circumstance G that, left unchecked, would have sidetracked the process that was leading to e. Of course, putting the effect in jeopardy is not all that c did, or it would not even resemble a cause. It also "saved" e by taking it out of the jeopardy. A c that threatens e with one hand (its G hand) while saving it with the other is certainly not a cause of e; but the effect does depend on it holding G fixed. This is illustrated by an example of Hartry Field's.[14]

> "Bomb": Billy plants a bomb under Suzy's chair; Suzy notices it and moves out of range before it has a chance to explode; a few hours later, she goes to a prearranged medical checkup, and is found to be healthy.

Let e be the event of the doctor's issuing a glowing report on Suzy's state of health. The glowing report is not caused by Billy's planting of the bomb. But it is in a certain (strained) sense *contingent* on Billy's action. If we hold fixed the fact that her chair exploded, Suzy would have been injured had she not been tipped off by the bomb. No doubt this is convoluted reasoning—one hardly wants to thank the bomb for alerting Suzy to a threat that it itself poses—but we have yet to see where it goes wrong.[15]

---

[12] They are called "enfeeblers" in "Advertisement for a Sketch of an Outline of a Proto-Theory of Causation."

[13] '>' is the counterfactual conditional operator; '$A > B$' says that, if it had not been that A, it would not have been that B. '$A >_Y B$' is in general short for '$(A \& Y) > B$'.

[14] The example dates back to a seminar that Field gave at CUNY Graduate Center in the early 1990s.

Here is a proposal, stated first in metaphorical terms and then gradually literalized. The problem with Billy's planting of the bomb is that it leaves us just as far from the effect as we would have been without it; "the road taken" is just as long as "the road not taken"—indeed, the one road includes the other. A little less metaphorically: everything the effect *would* have needed, had the bomb not been planted, it *continues* to need with it planted. One can cash the metaphor further by treating needs as events depended on under particular stated conditions.[16] *G* and *H* having been used for obtaining conditions (facts), let us introduce *K* for conditions that may or may not obtain.

> *K* puts *e* in need of *x* if and only if $\neg Ox >_K \neg Oe$, that is, *e* would not have occurred, had *x* not occurred in *K*-type circumstances.

If *F* is the (nonobtaining) condition of *c* failing to occur, and *H* is an appropriately chosen obtaining condition, then

> 'what the effect would have needed' = the events that *F* puts *e* in need of, and 'what the effect does need' = the events that *H* puts *e* in need of.

To say that 'Everything *e* would have needed, it continues to need' is to say that the second set subsumes the first, and with no help from *c*. What this means at an intuitive level is that the need for *c* is over and above what would, but for *c*, have been *all* the effect's needs. A need that is piled arbitrarily on top of what would otherwise have been all the needs is trumped up or artificial. The objection to *c* is thus very simple: it meets no real (nonartificial) need.

Now for some refinements and qualifications. History has a branching time structure, we assume. There is the trajectory actually taken through logical space; branching off of that are various alternative trajectories corresponding to other ways things could have developed. One branch in particular corresponds to the way things would have developed if c had not occurred. This, after it begins to branch off from actuality, is called the *fallback scenario*; the *actual scenario* is what happens after the branch-off point in actual fact. The events e depends on in the fallback scenario are the *fallback needs*:

---

[15] Hall—"Causation and the Price of Transitivity," this JOURNAL, XCVII, 4 (April 2000): 198-222—offers a detailed discussion of this sort of case.

[16] If it seems odd to think of events as needs, remember that 'need' can mean thing that is needed. ("The dogsled was piled high with our winter needs.") Needs in the ordinary sense do not exist in our system. Their work is done by events considered under counterpart relation *C*, the relation of meeting-the-same-need-as.

$$\text{FAN} = \{x \mid \neg Ox >_F \neg Oe\}$$

(remember that F says that c fails to occur).[17] The effect's *actual needs*, for a given choice of H, are the actual-scenario events that it depends on modulo H:

$$\text{HAN} = \{x \mid \neg Ox >_H \neg Oe\}^{18}$$

H makes the need for c artificial if and only if HAN covers FAN with c to spare; or, what comes to the same, FAN coincides with a c-free subset of HAN. The point either way is that the effect's actual needs c *apart* subsume its fallback needs. The would-be cause "makes itself indispensable" in the pejorative sense of puffing the effect's needs up beyond necessity, until finally it too can claim to have played a role.

So, what does the glowing medical report depend on, if Billy does not plant the bomb? Some pills have to be swallowed; Suzy must dodge the falling piano; the doctor's alarm clock has to go off; and so on. The problem with Billy's planting of the bomb is that it does not diminish these needs one iota. One does not even have to hold anything fixed; HAN contains all the fallback needs even if H is the null fact (the fact that always trivially obtains).[19] It also contains them on many other choices of H, including, ironically, the fact that put e in need of Billy's action in the first place: that Suzy's chair explodes. There are plenty of facts exposing the need for c as artificial.

IV. COUNTERPARTS

I said that the fallback needs (FAN) should *coincide* with a c-free subset of the actual needs (HAN). This has to be understood in a particular way. Suppose that Suzy gets sick if she does not stay hydrated. She sets her Palm Pilot to remind her to take H_2O at noon. Had Billy not planted the bomb, Suzy would have been in the chair at noon; the need would have been met by an act of water drinking, water being the one hydrous stuff available in the room. Since Billy did plant the bomb, Suzy was out on the sidewalk when the need arose; the need was met by eating Italian ice, that being the one hydrous stuff available outside. Assuming that eatings are essentially not drinkings, the event actually needed is not the one that would have been needed, had Billy not planted the bomb. But then *e* has fallback needs distinct from all of its actual needs, whence it would seem that that HAN does not cover FAN at all, let alone with *c* to spare.

---

[17] It might be wondered whether FAN contains the events it is supposed to: the events that e would have depended on had c not occurred. x is in FAN if and only if $(\neg Ox \,\&\, \neg Oc) > \neg Oe$, while it is $\neg Oc > (\neg Ox > \neg Oe)$ that defines would-be dependence on x The export-import law for counterfactuals—$P > (Q > R)$ if and only if $(P \,\&\, Q) > R$—brings the two requirements into alignment. I assume that the law is close enough to correct for our purposes, or anyway that the indicated consequence is close enough to correct.

[18] Remember that FAN and HAN are limited to events occurring *after* the branch point (the point at which the nearest c-less world peels off from actuality).

[19] It simplifies matters to pretend that there is only one such fact.

I answer that different *event* does not have to mean different *need*. Suzy's water-drinking and ice-eating meet the same need, as that need manifests itself in their respective scenarios.[20] The need for *c* is artificial if for each *x* needed in the fallback scenario, the effect's actual needs include, not perhaps that very event, but *an* event meeting the same need. Rather than each *x* in FAN having to be *identical* to a member of HAN other than *c*, each such *x* needs a *counterpart* in HAN, where counterparts are events meeting the same need.[21]

What does it mean for events to meet the same need? The idea is this. Fallback needs can be paired off with actual ones in ways that preserve salient features of the case: energy expended, distance traveled, time taken, place in the larger structure of needs. One wants to preserve as many of these features as possible, while finding matches for the largest number of fallback needs. One asks: How much of the fallback structure is embeddable in this way into the actual one? The maximal embedding determines which events meet the same need. (Should the actual structure succeed, with no help from *c*, in absorbing *all* of the fallback structure, the need for *c* is artificial.)

---

[20] Also, same event does not have to mean same need. An event that would have met one need in *c*'s absence might in the actual scenario meet another (or it might meet no need at all). It is better not to treat needs as separately existing entities, but suppose for a moment that we do: x_w is the need met by event *x* in world *w*. Then what is claimed in the text is that $x_@$ can be y_w even if $x \neq y$, and the claim of this note is that $x_@$ can fail to be y_w even if $x = y$.

[21] I shall be taking counterparthood to be symmetric and one-one. But there might be reasons for relaxing these requirements. Take first symmetry. There might be an *x* in FAN whose closest actual correspondent meets, not the same need as *x*, but an even larger need: a need with the need met by *x* as a part. This closest actual correspondent ought to qualify as a counterpart of *x*. So, the argument goes, counterparts should be events meeting *at least* the same need, which makes counterparthood asymmetric. The one-one requirement seems overstrict for a similar reason. It might take a pair of events to meet the need *x* meets all by itself in the fallback scenario; or vice versa. I propose to ignore these complexities.

V. THE DE FACTO THEORY

I have been downplaying the fact that artificiality is relative to a choice of *H*. But the relativity is important to our larger strategy. That *e* is put in need of *c* by an obtaining condition *G* testifies to the existence of a causal relation between *c* and *e*. Sometimes, though, the testimony misleads. What distinguishes the *G*s that do signal a causal relation from the ones that do not (the "deceptive" *G*s)? An idea that did *not* work was to credit all *G*s that were sufficiently natural. But a related idea held promise. That the need *c* meets is artificial by the lights of an obtaining condition *H* testifies to the *non*existence of a causal relation. Rather than ask *G* to be natural in itself, let us require it to be more natural than the *H*s, if any, testifying the other way.

So: one event *de facto depends* on another if and only if it is put in need of the other by a *G* more natural than *H*'s (if any) exposing the need as artificial. That can be taken as a definition. The proposal is that

(DF) *c* is a *cause* of *e* if and only if *e* de facto depends on *c*[22]

Later sections will compare (DF) to other recent theories of causation. Here I try it out on two well-known traditional challenges, *preemption* and *overdetermination*.

An example of preemption is given in our opening paragraph. How is it that Suzy's throw caused the window to break, when Billy threw only a moment later? The fact is that Billy's rock never comes into contact with the window; it never even gets close. That Billy's rock never gets close puts the effect in need of Suzy's throw. But is the need perhaps artificial, that is, superadded to what would otherwise have been all the effect's needs?

The effect's needs if Suzy had not thrown are for Billy's throw, his rock traveling toward the window, its hitting the glass, and so on. These would seem to recur in the actual situation as needs for Suzy's throw, her rock's traveling toward the window, and so on. But then Suzy's throw meets the same need as was met in the fallback scenario by Billy's throw. The need was thus preexisting; it would still have had to be met even if Suzy had elected not to throw. If someone thinks that the need met in the fallback scenario by Billy's throw does *not* recur in the actual situation, that is fine, too. Suzy's throw cannot be over and above the fallback needs unless those needs are still present; and on the suggested hypothesis, one fallback need, at least, is eliminated.

---

[22] I leave the usual qualifications to be understood; both events should occur, for instance, and each should be possible without the other.

This sort of pattern arises often enough to call for some new terminology. Say that c *meets* a fallback need if and only if it meets the same need as some x in FAN would have met in its absence; and that it *cancels* a fallback need if and only if there is an x in FAN such that *no* actual event meets the same need as x would have met in c's absence. It is easy to check that the need for c is artificial if and only if c does not address any fallback needs, meaning that it neither meets any fallback needs nor cancels any. Now, suppose it is known that c meets a certain fallback need if any actual event does. Then we can conclude at once that the need for c is not artificial. For if c does not meet the need, then nothing meets it; and if nothing meets it, then (given that the effect does occur), the need was cancelled.

Now we turn to the second traditional challenge. Overdetermination occurs when an effect depends on two events taken together, but not on either alone; and (what distinguishes it from preemption) neither of the two events can lay claim to being more of a cause than the other.

It might happen, for example, that Billy decides to throw a little earlier, so that his rock hits the window at the very moment that Suzy's does. Billy's throw is not a cause by the lights of the counterfactual theory, for the effect does not depend on it (we can suppose there are no chains of dependence, either). Suzy's throw is similarly disqualified. But then what does cause the window to break? Not the conjunction of the two throws, since the effect could too easily have occurred without it. Not the disjunction, because we are hard put to regard the disjunction as a genuine event.[23] Could it be that nothing causes the window to break? This goes somewhat against the grain. An event that was caused (the breaking was not a miracle!) should, one feels, have causes. I do not mean to be blaming our difficulties here on any particular theory; I agree with Lewis that the case is *intuitively* undecidable. The question is what to make of this datum. Lewis contends that overdetermination can be left as "spoils to the victor." But why not aim for a theory that shows a similar indecision to our own?

(DF) calls Suzy's throw a cause if and only if some G putting the effect in need of it is more natural than any H making the need artificial (likewise for Billy's throw). It is not hard to find suitable Gs. The shattering depends on Suzy's throw modulo the fact G_s that Billy's rock does not hit the window unaccompanied, and on Billy's throw modulo the fact G_b that Suzy's rock does not hit unaccompanied. It is not hard to find suitable *H*s either; indeed, we have already found them. G_s makes the need for Billy's throw artificial, and G_b does the same for Suzy's throw. For suppose that Billy had not thrown. The shattering would have depended on Suzy's throw, the forward motion of her rock, and the like. Holding fixed that Billy's rock does not hit alone (that is G_s), these events are still needed in the actual situation. Thus $G_s = H_b$, and for similar reasons, $G_b = H_s$.

---

[23] Lewis, "Events," in his *Philosophical Papers, Volume 2*, pp. 241-69; see especially section VIII, "Events Are Not Disjunctive."

Assuming that these are the most natural cause makers (breakers) to be had, is the effect put in need of Suzy's throw by a fact more natural than any fact making the need artificial?

That depends. One reading of 'more natural' is *strictly* more natural. If that is what is meant, then neither throw is a cause; each prima facie connection is broken by a fact exactly as natural as the one that established it. But the phrase could also be taken weakly, to mean 'at least as natural as'. It is true that each throw occurs under conditions modulo which the effect ignores it. But then each also occurs under conditions no less natural modulo which the effect needs it. Tie goes to the runner on the weak reading, so we have two bona fide causes. Our intuitive indecision about whether to treat overdeterminers as causes reflects semantic indecision about what to mean by 'more'.[24]

VI. THE COUNTERFACTUAL THEORY

Event *e* depends counterfactually on event *c* if and only if it would not have happened, had *c* not happened.[25] If we write R^* for the ancestral of a relation *R*—the relation *x* bears to *y* if and only if they are connected by an *R*-chain—then the counterfactual theory says that

(CF) *c* is a cause of *e* if and only if *e* depends* on *c*.

It is understood that *c* and *e* both occur, that they are suitably distinct, and that various unnamed other conditions are met. Before considering some differences between (CF) and (DF), I want to point out a respect in which they agree. If *e* depends on *c* (and the unnamed other conditions are met), then *c* is a cause of *e* according to both theories.

This is clear on the counterfactual theory because a relation entails its ancestral. But why does dependence entail causation on the de facto theory? If *e* depends on *c*, then the null fact (the fact that always trivially obtains) puts *e* in need of *c*. It now has to be shown that no comparably natural H exposes the need as artificial. But no H can expose the need as artificial. Artificiality presupposes fallback needs—events that e depends on in c's absence—and e does not even occur in c's absence.[26]

---

[24] This is intended more as a rational reconstruction of our intuitive response than a psychological explanation of it.

[25] "Dependence" is counterfactual unless otherwise noted.

Now let me list some features of the de facto theory that, depending on one's other commitments, might seem to recommend it.

Not all preemptive causes register with the counterfactual theory. Early or "cutting" preempters get through, but late preempters, as in our opening example, are reckoned ineffective. A dependence chain from Suzy's throw to the shattering cannot be found, because there is no point whereafter only her rock is positioned to break the window. I am not claiming that the de facto theory catches all preempters of whatever type; I have not seen all the types. It does, however, hold out hope of catching them all.

Second, (DF) is a "sine qua non" account; the cause is a that without which not. (CF) gives up on the sine qua non idea because of preemption. But this is to snub the intuitions that initially recommended the theory. As Judea Pearl says: "...why should the existence of a counterfactual dependence chain be taken as a defining test for a concept as crucial as 'actual cause', by which we decide the guilt or innocence of defendants in a court of law? The basic counterfactual criterion does embody a pragmatic rationale;...we would like to encourage people to watch for circumstances where their actions could make a substantial difference. Once the counterfactual dependence between the action and the consequence is destroyed..., what good is it to insist on intermediate counterfactual dependencies along a chain that connects them" (*op. cit.*, p. 313)? De facto-ists think that ancestralizing is an overreaction. Preemption can be dealt with another way.

The counterfactual theory makes causation out to be transitive. But recent years have seen a number of counterexamples to transitivity. Consider again "Bomb." Billy's putting a bomb under Suzy's chair was a cause of her moving away; her moving away was a cause of her later being pronounced healthy; but his putting the bomb under her chair was not a cause of her later being pronounced healthy. The de facto theory can take these examples at face value. That e de facto depends on a de facto dependent of c leaves the question of e's de facto dependence on c wide open.[27]

---

[26] Suppose that H does somehow manage to make the need artificial. That upsets the de facto dependence of e on c only if H is as natural as the null fact. Since dependence modulo the null fact entails dependence modulo every fact independent of c—every fact G such that $\neg Oc > G$—all such facts would have to be less natural than H before the de facto dependence claim was affected.

Even if some form of dependence is sufficient for being *a* cause, it is not sufficient for being *the* cause. An event c is not *the* cause of e unless it is proportional or commensurate with e. (This might be taken to mean that c is not screened off from e by its determinables, while its determinates are screened off from e by c itself.[28]) It is unclear, though, how the proportionality constraint can apply when e depends on c only indirectly. There are likely to be lots of dependency chains, assigning c lots of different successors. With which of these successors does c have to be proportional if it is to come out proportional to e? The de facto theory avoids this problem by insisting on one-off dependence.

VII THE POTENTIAL ANCESTOR THEORY

One strategy against preemption (mine) is to bring hidden dependence relations to the surface by holding certain things fixed. A different strategy tries to bring them to the surface by holding some things fixed while changing others.[29] What changes is that other candidate causes are taken away. What is held fixed is that no "new" events—none, anyway, on which the effect depends*—are allowed to occur. This works well for the broken-window example. Had Billy not thrown, the effect would have depended* only on Suzy's throw and other actual events. Had Suzy not thrown, however, the effect would have depended* on events that did not occur, like the window's being hit by Billy's rock.

But there seem to be other examples where the preempted "cause" would not need new events to generate e. Suppose we are playing the kind of card game where success depends on quick reactions. Suzy and Billy are on one team, you and I are on the other. The Queen of Hearts has just turned up, and the winning team is the one that touches it first. Suzy and Billy both lunge at the card, with Suzy's finger arriving a moment earlier. It was Suzy's lunge that ended the game, causing us to pound the table in frustration. If Suzy's lunge had not occurred, though, we would still have pounded the table as a result of events all of which actually occurred, prominent among these being Billy's lunge.

---

[27] There is no reasonable compounding operation on facts with the property that, if d depends modulo Q_1 on c, and e depends modulo Q_2 on d, then e may be expected to depend modulo $Q_2 \bullet Q_1$ on c. Any reasonable compounding operation has $Q \bullet Q = Q$, so it is enough to show that dependence modulo Q is not always transitive. But we know it is not when Q is the null fact; this is just the ordinary intransitivity of counterfactuals. For more on transitivity, see Michael McDermott, "Redundant Causation," *British Journal for the Philosophy of Science*, XL (1995): 523-44; and Hall, "Causation and the Price of Transitivity."

[28] See my "Cause and Essence," *Synthese*, XCIII (1992): 403-49, and "Mental Causation," *Philosophical Review*, CI (1992): 245-80

[29] Jonardon Ganeri, Paul Noordhof, and Murali Ramachandran, "For a (Revised) PCA-analysis," *Analysis*, LVIII (1998): 45-47

A second example is due to Alex Byrne and Ned Hall.[30] Suppose that *c*-type events produce *e*-type events directly, that is, with no help from intermediaries. The same is true of c'-type events, but it takes slightly longer: the effect follows in 1.1 seconds rather than 1 second exactly. If *c* and c' take place at noon and *e* occurs precisely 1 second later, then (other things equal), *c* is the preempting cause and c' is the preempted. But had *c* not occurred, *e* still would have, and without any dependence on nonactual events.

What does (DF) say? Consider first the card game. An obvious choice of *G* is the fact that over the whole course of the game, Billy never touches the Queen of Hearts. Holding that fixed, if not for Suzy there would have been no table pounding. In the Byrne/Hall example, we can let *G* be the fact that *e* does not occur more than a second after c'. Had *c* failed to occur in *G*-type circumstances, *e* would not have followed, for the slower-acting c' would not have had time to produce it. Unless the need can be made to look self-inflicted by an equally natural *H* (I cannot think of one), *e* de facto depends on *c* and so is one of its effects.

VIII. THE HASTENER THEORY

A third approach defines causes as events in whose absence *e* either would not have occurred at all, or would not have occurred so early.[31] If we take "never occurring" to be the limiting case of delay, the theory defines causes as events without which the effect would have been delayed. Since Suzy's lunge and faster-acting *c* make their respective effects happen earlier than they otherwise would have, the hastener theory can deal with the examples of the last section. But neither it nor the potential-ancestor theory can deal with Jonathan Shaffer's[32] trumping preemption:

> Imagine that it is a law of magic that the first spell cast on a given day [matches] the enchantment that midnight. Suppose that at noon Merlin casts a spell (the first that day) to turn the prince into a frog, that at 6.00pm Morgana casts a spell (the only other that day) to turn the prince into a frog, and that at midnight the prince becomes a frog. Clearly, Merlin's spell...is a cause of the prince's becoming a frog and Morgana's

---

[30] "Against the PCA-analysis," *Analysis*, LVIII (1998): 38-44

[31] Paul, "Keeping Track of the Time: Emending the Counterfactual Analysis of Causation," *Analysis*, LVIII (1998): 191-98

[32] "Trumping Preemption," this JOURNAL, XCVII, 4 (April 2000): 165-81.

> is not, because the laws say that the first spells are the consequential ones. Nevertheless, there is no counterfactual dependence of the prince's becoming a frog on Merlin's spell, because Morgana's spell is a dependency-breaking backup (*ibid.*, p. 165).

The potential ancestor theory is no help because "there is [no] failure of intermediary events along the Morgana process (we may dramatize this by stipulating that spells work directly, without any intermediaries)" (*ibid.*). As for the hastener theory, there is no "would-be difference in time or manner of the effect absent Merlin's spell" (*ibid.*).[33] It seems that "nothing remains by which extant [counterfactual accounts of causation] might distinguish Merlin's spell from Morgana's in causal status" (*op. cit.*).

De facto dependence is established by holding fixed that *only Merlin casts a first spell*, meaning by that that nobody else does. Had Merlin held back, then given that other spells, if any, wait on his, no spell would have been cast, hence the prince would have stayed one. Of course, the possibility remains that a no less natural H can be found that exposes the need as artificial. But as was argued in section v, such an H is at odds with the intuitive judgment that Merlin's spell meets the same need as Morgana's would have, if any actual event does. And an H flouting that judgment would have, one suspects, to be too unnatural to matter. But we need not settle this here; if (DF) is not at an absolute loss when it comes to trumping preemption, then it is at less of a loss than extant alternatives.

IX CAUSATION AND INFLUENCE

Most extant alternatives, anyway. A theory recently proposed by Lewis[34] can deal with all of the cases so far mentioned. Very roughly, Lewis holds that c causes e if and only if c has variants c', c'',...and e has variants e', e'',...such that which e-variant occurs depends on which c-variant occurs. (That is direct causation; causation is the ancestral.) Since the prince's fate is sensitive to the content of Merlin's spell but not Morgana's, it is Merlin's spell that does the causing. But consider a Merlin-type scenario due to Collins: "There is only one spell [Merlin] may cast (standard prince-to-frog) and only one time of day he may cast it (noon). No such limitations apply to Morgana.... The prince's transfiguration is now in no way dependent on...how Merlin acts.... Yet...it is still Merlin's spell, and not Morgana's, that causes the prince to turn into a frog."[35] This appears to show that a cause need not influence its effect, while a noncause can influence the effect greatly.

---

[33] For related difficulties, see Hugh Rice, "David Lewis's Awkward Cases of Redundant Causation," *Analysis*, LIX (1999): 157-64.

[34] "Causation as Influence," this JOURNAL, XCVII, 4 (April 2000): 182-97.

Our task is to find a *G* modulo which the transfiguration depends on Merlin's spell rather than Morgana's, despite that the influence is all on the side of the latter. The same *G* used above appears to work here as well. Had Merlin held back in *G*-ish conditions, there would have been no spells at all, and so no transfigurations. If there is a more natural *H* portraying the need as artificial, it is not easy to find.

How much influence must a preemptive cause exercise over how its effect occurs? None, if the how is a how of *manner*. But suppose that it is a how of *means*. One feels that an event *greatly* redirecting the process leading to *e* is a cause, while one *slightly* redirecting it is not.[36] Consider

> "Flip": A track splits briefly into subtracks A and B and then reconverges. Further on lies the station. Which subtrack a train takes to the reconvergence point is controlled by the position of a small electrical switch. Railroad worker Billy, bored with his job, has taken to flipping the switch back and forth when he hears a train approaching. Today's train would have used subtrack A had Billy's (final) act of switch-flipping not redirected it to B.

No doubt that final flip does cause the train to arrive by one route rather than the other. But does it contribute to the train's arriving as such? I take it the answer is that it does not. This might seem to create a problem for the DF theory, since the arrival depends on the final flip modulo a quite natural *G*. Holding fixed that nothing goes down subtrack A, there would have been no way forward had the switch not been flipped. Are we then not overgenerating, as Lewis undergenerates? His account demands a kind of influence that is not essential to causation; DF fails to demand a kind of influence that is essential.

But this is to forget an important part of the theory. It is true that *G* puts the effect in need of that final flip. But the need is easily exposed as artificial, for example, by the null fact. The effect's fallback needs—its needs if the final flip had not occurred—are, let us say, for 100 one-yard motions down A. Does Billy's flipping of the switch meet any of these needs? Clearly not. Does it cancel any? No; they persist into the actual scenario as needs for 100 one-yard motions down B. By the lights of the null fact, then, no real need is met by Billy's action but only an artificial one. (It might be thought that the null fact is no more natural than G. But if it makes a need artificial, then so do all "insensitive" facts: all facts H such that $\neg Ox > H$ whenever $\neg Ox > \neg Oe$.[37] G is sure to be less natural than some such fact.)

---

[35] "Preemptive Preemption," p. 231.

[36] This is a particular theme of Hall, "Causation and the Price of Transitivity."

Note that, if we change the example so that fallback needs are cancelled, our causal intuitions change too. Suppose that subtrack A is damaged as the train approaches. The repair crew is working madly to fix it; they cry out to Billy "Flip your switch!" and he does. Now Billy's action plays a definite role. It does not matter if subtrack A is by some miracle repaired just in time. By relieving the effect of its need for that miracle, Billy's action makes itself eligible for the role of cause.

This is a good place to acknowledge that, although technically one fallback need is as good as another, in practice not all such needs are taken equally seriously. What if subtrack A is *normally* disconnected? The drawbridge it runs over is kept open to let boats through, except when sensors pick up an approaching train, at which times it routinely and automatically closes. The need that Billy's action cancels is, let us say, the need for a generally reliable mechanism to work the way it is supposed to. Such a need may be considered too slight to protect the action that cancels it from charges of artificiality. I am not sure what makes a need "serious" enough to escape this sort of criticism. One key factor seems to be this: how far-fetched the notion is of its not being met.[38] Far-fetchedness is a largely objective matter, to do with the prior probability of the need's not being met, and the remoteness of the scenarios where that happens. But nonobjective factors may play a role, too, such as how worried we are about the effect's not occurring for that particular reason, and how impressed we are by the size of the canceled need.[39]

---

[37] It is enough to show that NAN $\subseteq$ HAN, that is, every event that the null fact puts e in need of, H puts e in need of as well. Suppose that x belongs to NAN. Then $\neg Ox > \neg Oe$, which since H is insensitive means that $\neg Ox > H$. The last two counterfactuals imply that $(\neg Ox \,\&\, H) > \neg Oe$, whence x belongs to HAN. (As usual, x ranges over events occurring after the point at which the nearest c-free world branches off from actuality.)

[38] I take the notion of far-fetchedness from Collins, "Preemptive Preemption."

[39] This bears on an example of McDermott's. "Suppose that I reach out and catch a passing cricket ball. The next thing along in the ball's direction of motion was a solid brick wall. Beyond that was a window. Did my action prevent the ball from hitting the window? (Did it cause the ball to *not* hit the window?)" —"Redundant Causation," p. 525. Reactions to this change as interest rises in the need that gets canceled: the need would have been met by the ball's hitting the wall. (I assume that the need is canceled; for if it recurs it is met by the catch, whence the catch is a cause. It may help to replace the wall with a powerful steady wind.) Usually, we think, 'The wall was all set to stop that ball; it could

I spoke of the feeling that a *c* making "big" changes in the process leading to *e* is better qualified for the role of cause than a *c* making "small" changes in that process. If small changes are changes leaving the fallback needs more or less in place, the de facto theory bears this feeling out. The following is an example of Ned Hall's (somewhat abridged).

> "Kiss": Billy and Suzy meet for coffee. Suzy kisses Billy passionately, confessing that she is in love with him. Billy is thrilled—for he has long been secretly in love with Suzy, as well. Much later, as he is giddily walking home, he whistles a certain tune. What would have happened had she not kissed him? Well, they would have had their usual pleasant coffee together, and afterward he would have taken care of various errands, and it just so happens that in one of the stores he would have visited, he would have heard that very tune, and it would have stuck in his head, and consequently he would have whistled it on his way home...even though there is the failure of counterfactual dependence typical of switching cases (if Suzy had not kissed Billy, he still would have whistled), there is, of course, no question whatsoever that as things stand, the kiss is among the causes of the whistling.[40]

That seems right: the kiss is a cause of the whistling. But the example is not really typical of switching cases, for the kiss makes major changes in the process by which Billy comes to be whistling. The whistling's needs in the fallback (kissless) scenario are weighted toward the period *after* Billy leaves the coffee shop. They include Billy's deciding to drop into that particular store, the store's staying open until he arrives, the radio's playing that particular tune, and so on. All of these needs, and they are not trivial, are cancelled when Suzy gives Billy the kiss.

Suppose I am right that it is because Suzy's kiss takes the whistling out of a certain kind of late-afternoon jeopardy that we are willing to credit it as a cause. Someone might say: the kiss also puts the whistling into *new* jeopardy; it now becomes important, for instance, that Billy steers clear of a certain malicious gossip. Why is everyday causal

---

and would have done so easily; the catch gets no credit whatever for relieving the effect of so piddling a need'. Occasionally, we think 'The ball had a great deal of momentum; the job of redirecting that momentum is not made any smaller by the fact that the wall was more than up to it; kudos to the catch for making the job unnecessary'.

[40] "Causation and the Price of Transitivity."

thought not more bothered by that? If the role of a cause is to minimize needs, should not introducing new needs be just as bad as preserving old ones?

The answer to this is that the role of a cause is *not* to minimize needs; it is to meet needs. It is just that the need has to be of the right type: real rather than artificial. Taking the effect out of jeopardy is what *c* does, not to fulfill its mission as a cause, but to establish the bona fides of a particular need. That having been done, it is meeting the need that makes *c* a cause.

STEPHEN YABLO

Massachusetts Institute of Technology

BOOK REVIEWS

Kant's Ethical Thought. ALLEN WOOD. New York: Cambridge University Press, 1999. xxiv + 436 p. Cloth $59.95, paper $21.95.

A strong believer in the ideals of the Enlightenment, Allen Wood has written this outstanding book out of a sense that in recent years our Enlightenment heritage has been severely eroded. What is needed, Wood holds, is a clear reaffirmation of these values, but one that is open-eyed to the bleak social and political realities of today; and it is Immanuel Kant's ethical thought, he maintains, that can provide especially useful guidance in this regard. There is nothing surprising, of course, in hearing the name of Kant mentioned in an appeal to Enlightenment ideals. But it certainly runs against a familiar picture of Kant to suggest that his thinking is in touch with the grimmer aspects of practical reality in a way that keeps its exalted ideals in meaningful connection with them. A chief aim of Wood's book, then, is to free our view of Kant from the misconceptions that prevent us from appreciating how the approach to ethics represented in his philosophy has pertinence today.

In pursuit of this aim, Wood advances two main contentions, which he elaborates in the two parts into which his book is divided. In the first part, which deals with the "metaphysical foundations" of Kant's ethical theory, Wood rejects a common view of Kant's ethical thought as "a metaphysical system of mindless rule-following," blind to the consequences of action; Wood maintains to the contrary that Kant's theory is in fact based on a robust conception of value, provided by the idea of the dignity of humanity, the rational nature equally present in all human beings. In the second part, Wood argues effectively that in his application of this theory Kant is guided by rich empirical conceptions of human nature, society, and history; these conceptions are open to criticism in some respects, but as they incorporate a deep Rousseauean appreciation of the human propensity to delude oneself with insupportable opinions of one's superior self-worth, a propensity that is rooted in our social nature and influences the trajectory of human history, an understanding of these conceptions enables us to see how Kant's ethical thought bears in a socially and historically informed way on the great inequalities of wealth and status that pervade modern civilization.

The first part traverses familiar terrain. Its five chapters almost amount to a commentary on the *Groundwork of the Metaphysics of*

0022-362X/02/9903/149–52

Morals, and disputed topics relating to the good will, duty, maxims, imperatives, and so forth all come in for extended and expert discussion. The problem at the center of Wood's attention arises from the fact that Kant presents his fundamental principle, the categorical imperative, in several versions, or "formulas," which seem plainly to be quite different in character and divergent in their implications. In addition to the vexed formula of universal law, which enjoins us to act only on maxims we can will to be universal laws, Kant presents two other, more favorably received formulas, those of humanity and autonomy, which introduce, respectively, the value of humanity as an end in itself and the idea that each bearer of humanity is vested with equal authority as co-legislator of the laws to which it is subject. Like some other recent exponents of the Kantian approach, such as Thomas Hill and Christine Korsgaard, Wood responds to this problem by organizing his reading of Kant's ethics around the formula of humanity, rather than around the more controversial formula of universal law. But Wood goes further than most in the extent to which he develops his criticism of the formula of universal law. This formula, he maintains, is only a partial and incomplete expression of the idea of the moral law, an idea that cannot be properly grasped without an appreciation of the fundamental ideas of the value of humanity and of autonomy. Rather than abandon Kant's theory on account of the problems surrounding this first formula, Wood advises, readers should push onward to the second and third, where the heart of the theory is to be found.

According to Wood, Kant errs in claiming that the formula of universal law can be derived from the mere concept of a categorical imperative. Wood grants that this concept does imply that the moral laws to which our maxims should conform hold universally for all rational beings, but he maintains that Kant confuses this legitimate implication with the fundamentally flawed thought that this concept entails that we must be able to *will* our maxims as universal laws. Moreover, as has been amply demonstrated in the literature, when we apply this formula we find that it not only fails to exclude maxims that are clearly impermissible, but also fails to accommodate maxims that are obviously unexceptionable. For Wood, then, the element of truth in the first formula is simply the minimal idea that moral laws, whatever they may turn out to require, are universal. It is not until we reach the fundamental conception of value provided by the later formulas that we are in a position to specify what the fundamental universal law of morality is and to find a suitable way of understanding Kant's talk of willing a maxim as a universal law. Thus, Wood suggests that whereas Kant's claim that the maxim of refusing to help others cannot be willed as a universal law depends on "some possibly controversial empirical assumptions about what is rational given the facts of human vulnerability and interdependence," assumptions that might be rejected by a rugged individualist, a less questionable argument becomes available once we draw on the second formula's conception of value and recognize that the reason why this maxim cannot be willed as a universal law is that refusal to help shows contempt for humanity (150).

Readers who are less critical of the first formula may doubt, however, that the formula of humanity provides a more fundamental understanding of what is wrong with this maxim. They may reply that, if it seems easy to imagine (Kant's claim to the contrary notwithstanding) that the rugged individualist can will to forgo others' help, then it seems equally easy to suppose that this person need not regard others' refusal to help as displaying contempt for humanity. Wood acknowledges that the application of the formula of humanity depends on intermediate empirical premises that specify what does and what does not count as respecting humanity. But the empirical premise that refusing to help others counts as a failure to respect humanity, or rational nature, seems itself to involve empirical assumptions about what is rational given the facts of human vulnerability and interdependence—the same assumptions as the ones on which Kant relies in claiming that the maxim of refusing to help cannot be willed as a universal law. If so, then the idea of a maxim that can be willed as a universal law and the idea of respecting humanity may be more closely linked than Wood suggests.

But even those who do not share Wood's understanding of how the different formulas are related can appreciate his excellent treatment of Kant's anthropological and historical views and his illuminating demonstration of the importance they have for a proper understanding of Kant's moral theory. It is not possible in this review to convey adequately the wealth of topics in the book's second part on which light is thrown by Wood's well-informed and perceptive treatment. But a sense of the breadth of its scope can be gained by noting that the four chapters that belong to it are devoted to Kant's conception of the science of anthropology and its teleological basis, to his conception of history, to the inclinations, the affections, and the social basis of the passions, and to the historical self-understanding of Kant's ethical thought. Ranging far beyond the body of work to which interpreters of Kant's ethics have for the most part confined their attention, Wood draws extensively from Kant's writings on history, politics, and religion, as well as from his lectures in anthropology, and many useful comparisons are made with other thinkers, such as

Jean-Jacques Rousseau, Johann Gottfried Herder, G. W. F. Hegel, and Karl Marx. The Kant that Wood portrays in these pages is a student of human nature who is at once both a close and shrewd observer of human psychology and social relations and also a philosophical historian with a grand and elaborate view of the development of the human race and human society.

Part of Wood's aim in examining the anthropological applications of Kant's theory is to expose certain misconceptions that arise from considering Kant's more theoretical ethical writings in isolation (for example, that Kant's theory is objectionably "individualistic"), and to show that many of the harsher elements in Kant's view—including not only specific moral opinions that we today find unacceptable, such as his attitudes concerning sex and race, but also more general doctrines, such as his view of the opposition between reason and natural appetite—trace to his empirical anthropology rather than to the moral theory itself. But Wood's central positive theme is Kant's conception of the competitive, unsociable element within human sociability itself and the role it plays in human history, the idea that the means by which nature promotes the end of culture, or the development of the natural capacities of human beings, and ultimately the highest ends of morality and freedom, is the human desire for others' esteem, which, being the source of discontent and antagonism, spurs us to action that indirectly furthers these ends; it is through this strife-ridden contribution of nature to the development of the human species that human rational capacities are cultivated to the point where human beings can themselves frame a conception of their highest ends and freely join together in the cooperative pursuit of them. As Wood sums up the idea: "Kant's ethical thought is fundamentally about the human race's collective, historical struggle to develop its rational faculties and then through them to combat the radical propensity to evil that alone made their development possible" (296). Wood elaborates this theme with great insight and moral-psychological subtlety, and in doing so exposes to view an often overlooked vein that runs very deep in Kant's thought, an idea that lies at the center of his vision of how, through the workings of nature and freedom, moral ideas are alive in human affairs.

STEPHEN ENGSTROM

University of Pittsburgh

On the Emotions. RICHARD WOLLHEIM. New Haven: Yale University Press, 1999. xiii + 269 p. Cloth $30.00.

Richard Wollheim's main concern in this intense and intellectually concentrated book is to "repsychologize" the concept of emotion. Rejecting the idea of a "one-one correspondence between the linguistic forms that we use to ascribe mental phenomena and the ways in which such phenomena ought to be classified" (20), Wollheim insists on the psychological reality of the emotions, which he identifies as mental dispositions with a characteristic role and history. (His discussion is thus an exercise in "applied philosophy," which does not limit itself to conceptual analysis but appeals additionally to "observation and experiment, to common usage and traditional lore," and to the "wisdom of the ages" (xi).) The role of emotion is to provide us with an "orientation" to the world of which belief presents a picture, and within which desire sets objectives to aim at. As to its characteristic history, Wollheim sees emotion as flowing from an attitude that develops in us toward something we identify—rightly or wrongly, actually or prospectively—as responsible for the satisfaction or frustration of one of our desires. Once formed, an emotion tends to find expression in behavior and to induce further mental dispositions, desires included; it can therefore generate action, but only indirectly, for "what directly generates action, here as elsewhere, is the motivating conjunction of desire and instrumental belief" (16). Nor should emotion be identified with (or held to entail) any evaluative judgment, for it seems to be a fact of human nature that we can experience (say) anger, contempt, or love without necessarily believing these responses to be objectively justified: "In our emotional life, we do not always feel ourselves to have right on our side" (115).

Wollheim applies this general account to the "moral emotions" (meaning, centrally, shame and guilt) by arguing that in these cases the attitude that constitutes the emotion is an attitude toward oneself "as a person." Such attitudes arise in response to changes, and especially to a falling-off, in the sense of security we derive from being in good standing with others. But while we might be tempted to assume that real-life "others" were the key sources of criticism and reproof in the moral consciousness of any given individual, Wollheim proposes that we turn this assumption on its head and treat as primary the case in which the criticizing agency is "non-real" or fantasized (167)—a figure set in place by processes of "introjection" and "projection."

One striking, and plausible, claim advanced in the course of this discussion is that the moral emotions are inherently liable to "heteronomy"—that is, that the opinions or standards implicit in them can differ from our own, and yet can still appear to us as authoritative. "[W]hen the person is made aware that he has fallen short of the required standards, and is thus held in less regard, he is led to see himself as he would if he accepted these standards, which he may well not. We may say that the authority regulates the person's attitude without necessarily touching his opinions" (153). For example, in a scenario borrowed from Max Scheler to illustrate the workings of shame, an artist's model becomes aware that she is an object of sexual interest to the man who is drawing her (159-63). She may take pleasure in this awareness, but she also feels shamed by it—a response Wollheim attributes to her sense of being under the eye of an imaginary (or, more accurately, a fantasized) observer, who will "hold her in lesser regard" on account of the different way in which the artist now views her. Yet "what is not necessary for the onset of shame is that the model should concur with the opinion that the observer has of her.... All that is necessary is that she should be aware of how she appears in the observer's eyes, that his eyes should make her aware of this, and that this should in turn make her anxious," in which case "it is for her as if, as if for practical purposes, she concurred with the observer's opinion of her" (161). The autonomy of the moral emotions, Wollheim argues—in other words, that state of character in which it is only a fall in *our own* regard, by reference to standards of our own, that can trigger these emotions—is "an achievement, due to much psychological work on our part, and is not originary" (152); it may even be to a large extent chimerical.

Assuming for the sake of argument, however, that the model does indeed reject the judgment passed on her by the imagined critic, and that her anxiety is not to be interpreted as the effect of a genuinely epistemic uncertainty about the pros and cons of what she is doing, we may feel that Wollheim underplays the potential brilliance of the achievement in question here. One could tell a story in which the conquest of the kind of "false shame," as Scheler's model might put it to herself, that once afflicted her regardless of her own opinions was an important theme in her personal formation (compare the exaggerated unease about "making an exhibition of oneself" that inhibits the public functioning of many people, not least of women). By contrast, Wollheim's emphasis on the role of fantasy in the construction of the criticizing agency promotes a vision of the "inescapability" of shame and guilt (188), and issues in a somewhat static account of the process by which a reflectively acceptable morality can (with luck) be arrived at: "The person projects [the criticizing agency] on to, or into, the environment in order, for instance, to come to terms with those in authority, or to find a measure of shared agreement with friends and acquaintances" (219).

A more serious worry is that Wollheim's elucidation of the concept of emotion in terms of the satisfaction or frustration of desire leads in places to some curiously reductive consequences, at any rate in regard to the negative emotions. Some mental gymnastics are required, for instance, in order to convince oneself that a boy who "wakes up in damp grass, finds a frog sitting on his chest, and develops a fear of frogs" will "perceive the frog as cause *at once of the frustration of his desire for tranquillity* and of the butterflies in his stomach" (121, 122; emphasis added). Tranquillity is undoubtedly a desirable condition, but in the absence of commotion, is there (psychologically speaking) any such thing as the desire for it? (If I unexpectedly tread on a slug with my bare foot and recoil in disgust, is this an instance of the frustration of my desire to avoid contact with the slimy?) The same oddity attaches to Wollheim's account of the "liberal who fervently desires that none of his fellow-citizens should suffer unnecessarily," and of whom we are told: "[O]ne day he learns that an acquaintance of his has been picked up and savagely tortured. A deep desire of his has been frustrated..." (71). Again, this characterization of the horror with which one would receive such news seems lame: the event does not so much frustrate one's desires as affront one's values.

Wollheim's inventory of negative emotions (61) also contains some phenomenologically puzzling entries such as "ingratitude," "indifference," and "hardness of heart." These are all dispositional properties, but it is unclear that any of them exhibits the structure of an unfavorable attitude to any particular object. Nor do they seem to originate in the perception of any object as frustrating, or threatening to frustrate, some desire on the part of their bearer (unless perhaps a hard-hearted person is supposed to be one who dislikes the inconvenience of being asked to act charitably or compassionately, and deals with it by suppressing in advance any inclination to comply—but this would need spelling out). Such properties seem rather to fall under the heading of a reprehensible absence of emotional engagement.

This is not an easy book to read. It is divided into three long chapters corresponding—though in a "massively enlarged" form—to the lectures on which it is based (the Ernst Cassirer lectures which Wollheim delivered at Yale in 1991). Each "lecture" is in turn divided into numbered sections, varying in length from a few lines to as many as twenty-three pages (though in the latter case some editorial error must be involved: Lecture I has two "section 8"s, including the twenty-three-pager, but no section 9 or 10; Lecture III, likewise, has two "section 4"s, but no section 5 or 12). But there is no analytical table of contents like the one included in *Art and its Objects*,[1] and Wollheim's complex array of theses, objections, intellectual diagnoses and replies, with their intricately interlacing strands of informal enumeration, is often hard to survey without such help. The main text is also purged of references to specific authors (apart from canonical names like those of Sigmund Freud, Melanie Klein, Stendhal, and Jean-Jacques Rousseau), and while many contemporary sources are named in the endnotes, this policy does nothing to diminish the density of the exposition. What is commendable, though, in Wollheim's writing is a certain stylistic austerity or restraint which seems to express confidence in the power of his material to speak for itself.

SABINA LOVIBOND

Worcester College/Oxford

---

[1] New York: Cambridge, 1980, second edition.

Truth and Hope. PETER GEACH. Notre Dame: University of Notre Dame Press, 2001. vii + 103 p. Cloth $24.00.

The appearance of a book by Peter Geach is an occasion of note, particularly as he is now in his eighty-sixth year. The expectation of interesting thoughts well expressed will not be disappointed, though the volume is slim and contains little that is really new. Most of the chapters were delivered as a series of lectures at the Internationale Akademie für Philosophie in Liechtenstein; but the book is, in effect, a short collection of loosely related essays, several incorporating passages from previous publications. The abiding impression is of the distinctiveness of Geach's philosophical writing.

Among thinkers in the analytical tradition, Geach is without peer as a literary stylist. One may disagree with his judgments, and take exception to the vituperative terms in which they are sometimes expressed; yet his writing is a model of clear and precise English, devoid of pseudo-subtlety but animated by a learned literary spirit.

Like Samuel Johnson and G. K. Chesterton, Geach is a pleasure to read; unlike them, he commands both plain prose and the technical languages of Greek and medieval philosophy, as well as being a powerful dialectician. Geach has made light of his claim to be a creative logician; yet his contributions to the philosophy of logic are rightly admired, and his deployment of logical insights in the course of nontechnical argumentation (particularly in ethics and philosophical theology) is without parallel in its lightness and efficacy. Earlier examples of the latter are his short essays on "Good and Evil" (1956),[1] and on "Ascriptivism" (1960)[2] which helped, in turn, to revive ethical naturalism and to all but put paid to expressivism.

The lightness of his exercise of logical acumen is also evident in his habit of making a profound point in the course of discussing other matters. More than once what he has briefly observed have later become objects of entire research programs. Two little-noted examples deserve to be recorded here. First, in his 1951 review of Gottlob Frege's *Grundlagen des Arithmetik*,[3] Geach considers the criterion of identity for numbers as objects. Noting Frege's analysis of 'There are just as many *F*s as *G*s' as 'There is one-one correspondence of the *F*s to the *G*s', he objects to Frege's treating identity as a relation between objects, but shows how the establishment of one-one correspondence without use of identity can be achieved with only minor amendments to the *Grundlagen* (though not so the *Grundgesetze der Arithmetik*). He then comments "Frege generates the infinite series of finite numbers by appealing to the fact that the finite number n is always immediately followed by the number of numbers from 0 to n inclusive. I think this procedure is valid" (*ibid.*, p. 541). The proposal that the derivation in the *Grundlagen* (resting on "Hume's Principle") is in order was subsequently advanced by Charles Parsons a decade and a half later, and then, in the 1980s, became the subject of important publications by George Boolos and Crispin Wright.

---

[1] *Analysis*, XVII, 2 (1956) 33–42.

[2] *Philosophical Review*, LXIX, 2 (1960): 221–25

[3] "Frege's *Grundlagen*," *The Philosophical Review*, LX, 4 (October 1951): 535–44.

Second, in his essay "On Worshipping the Right God" (written circa 1960 and later published in *God and the Soul*[4]), Geach introduces examples to show the necessity of distinguishing between what he there describes as "personal and impersonal uses of the verb 'to refer'." The examples run in two directions: some showing that a description may be satisfied by an individual though that is not the person to whom the speaker was referring; others that the speaker achieves reference though the definite description he offers fails to be satisfied by the referent. It was some years later that Keith Donnellan drew attention to attributive and referential uses of descriptions, and showed that reference may be achieved even though descriptive fit is absent.

There is nothing, I think, of comparable significance in the present essays, though there are two discussions that deserve serious attention and invite further development. These come in chapters 4 and 5: "Truth, Truthfulness and Trust" and "On Truth as an End" (the two most closely related essays). In the first, Geach argues that reliance on testimony and on authority, as on memory, are not warranted by induction from experience. Skepticism concerning them has to be exceptional and can operate only against a general background in which they are accepted as reliable. The fact that we must depend on these sources, if we are to develop and extend our knowledge, provides the basis for an account of the character and gravity of the wrongness involved in breaches of trust and in lying.

For Geach, the duty of not lying does not in general derive from a potential auditor's right to be informed; rather, it is to be understood teleologically in terms of per se tendencies toward the bad, including the corruption of one's mind, as well as the damage done to the institution of language through the debasement of its currency, and to the common enterprise of building lives upon truth. On this account, while all lying is bad it is not all equally so, for the tendencies involved may, I take it, vary according to intent, circumstance, and effect. Geach is keen to avoid any suggestion of consequentialism, but I doubt that he would deny that effect is a factor relevant to the evaluation of particular acts of intended falsity.

---

[4] New York: Schocken, 1969.

So far as promises are concerned, his view is that they are expressions of intention which entitle recipients to form legitimate expectations upon which they themselves can then act. A change of intention need not reveal the original expression to have been deceitful. Accordingly, the wrongness of breaking a promise is not per se that of a lie but is a matter of the benefit lost or the harm inflicted. This latter has also to take in the fact that the keeping of a promise might involve doing something of a kind that is absolutely forbidden, in which case whatever other bad results might be expected to ensue, the promise is no longer binding. Thus the conclusion emerges that the wrongness and gravity of breaking a promise are much more highly contingent than are those of lying.

In "On Truth as an End," Geach reprimands Stephen Stitch and Jane Heal for holding that truth has no intrinsic value, and then sets about showing how it can and should be our goal. Having dismissed pragmatist, performative, and correspondence accounts, he advances Thomas Aquinas's "conformative" theory (my term) according to which the truth of a judgment by *A* that *x* is *f* consists in the co-occurence of one and the same form in the object and in the mind of *A*. The problem, as Geach sees it, is how to extend this approach so as to construct a comprehensive account of the truth of judgments, including negative propositions such as is put forward by *A* in judging that *x* is not *f*. By way of an attempt to get clearer about truth, Geach then considers the syntactical and semantical character of *duality* as implemented for a natural language such as English. Self-standing sentences are dual to their contradictory opposites in the derived language; names are self-dual, as is negation; but predicables and terms for relations are dual to their contradictories; and quantifiers are different again.

The task of translation across the language is neither trivial nor obvious. What Geach arrives at, however, is the conclusion that the unit of discourse apt to be the content of a judgment is always dual to its negation. The distinctive semantic role of judgments (positive and negative) is then understood in terms of their orientation toward truth. The metaphors of alignment and of direction allow for the idea that a false judgment is still in alignment with the true even though it is pointing away from it. We do not need to suppose that there is a corresponding object "the false" to which it is directed: truth alone will suffice for understanding the true and the false. Geach then goes on, somewhat in the manner of Anselm, to identify "Truth" with God (in *De Veritate*,[5] Anselm concludes that "the highest truth subsists in itself and belongs to no thing; but when something is in accord with it, we then speak of its truth or rectitude" (*ibid.*, p. 174)). Geach allows that this identification involves a further step; but even those disinclined to follow him that far should find much of interest in the earlier discussion of truth and duality.

Like *God and the Soul*, Geach's first collection of essays, *Truth and Hope* ranges across the fields of metaphysics, philosophy of mind, philosophy of religion, and the philosophy of virtue. For the most part, what he has to say is, in both style and substance, congruent with earlier writings. There is, however, one important issue on which there is some tension between what he now insists upon and what he previously advanced in equally forcible manner. In "What Is Man?" (chapter 2), Geach argues, as he has done before, that man's rational nature cannot be explained in natural terms, that is, as arising from matter through a series of physical, chemical, and biological processes. In *The Virtues*,[6] he followed W. V. Quine in emphasizing the logical difference between the propositions of natural science, and the intentional idiom required for the description of thought, language, and action; though in contrast to Quine, he insisted that psychological discourse was truth bearing. Here, he goes further in detailing the range of rationality and in arguing, in particular, that the practice of inductive reasoning cannot be explained, let alone justified, in Darwinian terms. His conclusion is that our rationality is naturalistically unaccounted for, but that it is explained on the assumption that we are the creation of an intelligent designer: "living images of a Divine Providence which controls not only the humanly observed course of events but also the movements of men's thoughts" (30).

Setting aside concerns about human freedom and divine foreknowledge (to which Geach turns in the interesting chapter on "Prophecy") I was struck by the resolute manner in which he dismisses doubts about the possibility of naturalistic explanations of the emergence of life: "I do not in the least wish to rule out either the origination of new life from non-living materials, or the development of new species"(26). For one thing, it might be supposed that biological phenomena per se exhibit teleological features analogous to intentionality, which are no less resistant to naturalistic explanation or reduction and which point to intelligent design. For another, Geach himself seemed once to argue that this is so. In *Providence and Evil*,[7] he maintained that the appearances of teleology which we observe among animals and plants are neither illusory, nor explicable by natural selection. For *species* to have originated there would have to have been a method of generational reproduction; but that capacity and the organs in which it is embodied are themselves ineliminably teleological phenomena. He then continued: "[explanation by natural selection being unavailable in this case] why then need we think some such story must account for other ostensibly teleological structures and activities of living things?... The spectacle of the living world, to an uncaptive mind, is a manifestation of great power and of great wisdom, for which no detail is too small" (*ibid.*, p. 77).

---

[5] *On Truth*, R. M. McInerny, trans., in B. Davies and G. R. Evans, eds., *Anselm of Canterbury: The Major Works* (New York: Oxford, 1998), pp. 151–74.

[6] New York: Cambridge, 1977.

Certainly, that earlier argument concerned the origination of species per se and not the derivation of one species *from* another; also in *Providence and Evil*, he does not mention the relation of life to inanimate matter. Yet the particular problem he there identifies for explaining away the teleological *as such*, encourages in him, as in others, the thought that its occurrence elsewhere is also due to design. Perhaps there are those who think that the locus of the philosophical naturalist's problem lies solely in the idea of natural *descent*; but if we are allowed teleology, then modification poses no special problem. If, however, teleology is not wholly explicable other than by reference to design, then this issue arises at the initial transition from the inanimate to the animate, for it is there that the behavior of entities first calls for explanations in terms of final causality. The problem of the emergence of life may not after all be easily distinguishable from the problem of the origin of species.

Geach is in general more likely to cite figures from the past than from the present; and when his contemporaries are mentioned it is often, with the major exception of Quine, in critical voice. At one point in the present book he names a notable philosopher of religion without mentioning that he has done much to revive the subject in recent times, but citing one of his works as illustrating the sort of thing that gets written when people are ignorant of the nature of consistency proofs. Geach's quite specific points are well taken—that a demonstration of consistency may be most easily given by establishing that a conclusion is soundly derived; and that imagination is an unreliable method for establishing coherence—but his dismissive style of expression is not to be emulated. Amusing as they may be, invective and ridicule are more likely to do harm than good in the context of writing about the sincere efforts of other undoubtedly able and well-intentioned philosophers. I mention this in conclusion because among Geach's admirers, of whom I count myself one, there is a tendency to adopt the manner in the hope of achieving the substance. In general, that would be self-deluding, and in cases of criticism of others it could well be uncharitable. Geach is a very considerable philosopher and a wonderful stylist; all the better had he been inspired by a gentler muse.

---

[7] New York: Cambridge, 1977.

JOHN HALDANE

Georgetown University
and
St. Andrews University

Emotion and Peace of Mind: From Stoic Agitation to Christian Temptation. RICHARD SORABJI. New York: Oxford University Press, 2000. x + 499 p. Cloth $45.00.

This volume has the wide scope and accessibility that readers of Richard Sorabji's books have come to expect. The subtitle actually gives too narrow a view of its brief, which is to discuss philosophical treatments of emotion from the Presocratics through late antiquity and in our Anglophone contemporaries, as well as in the recent brain research of Antonio Damasio and Joseph LeDoux. Besides the core topic of the Stoic theory of emotion, we find subjects as diverse as emotion and music, catharsis, the will, and the afterlife of Stoic materials in the form of the seven cardinal sins.

While descriptions claiming that some effort caters to philosophers and classicists, students and experts alike are now too commonplace to be effective, this book really lives up to that billing. Anyone who has some interest in any part of this history and would like to connect it with any other part will find the book useful. The references to an incredible range of primary and secondary works together with reports about their contents are of considerable value, and there are not many scholars who could even have attempted to provide such a collection. For this Herculean achievement, Sorabji deserves our gratitude.

Let me now turn to another aspect of the book's purpose. As Sorabji charmingly puts it, "Stoicism can be very helpful in dealing with counter-productive emotion. But it is not a matter of gritting your teeth. It is about seeing things differently, so that you do not need to grit your teeth" (1). The introduction in fact indicates a dialectical project of using the Stoic work on emotions in a novel way. The idea is as follows. The Stoic Chrysippus not only expounded the view that emotions are judgments, but specified the judgments in question. In Sorabji's view of this analysis, each emotion is constituted by a pair of judgments: one to the effect that something (which is in fact indifferent) is good or bad, the other that a certain response is appropriate. Thus fear is the conjoined beliefs that something bad is coming my way and that it is appropriate to avoid it, while pleasure is the belief that I have got something good with the additional judgment that internal felt expansion is an appropriate response. Since the Stoics held that every emotion involves commitment to a false belief, they were led to their famous demand for the extirpation of emotions. Emotions, as judgments, require assent if they are to come into being, and Stoicism holds itself out to help a person question appearances and avoid assent, thus avoiding emotion.

Sorabji recommends that we adapt this Chrysippan therapy for our own use. He wants us to do so, however, without signing up to become card-carrying Stoics. On two points he regards the orthodox view as being mistaken. First, he rejects the peculiarly Stoic view that nothing but virtue matters, which is foundational to the rejection of all emotion as involving false beliefs. And second, he argues that Chrysippus's analysis of emotions as judgments is not actually correct. Let me take up each of these points in turn.

On the first point, Sorabji writes: "We do not have to agree with this unacceptable side of Stoicism in order to learn from the Stoics how to be free of those emotions that are unwanted or counter-productive" (7). Clearly, this strategy will work to the extent that the Stoics have provided therapy that is independent of their general views on value. The proleptic remarks of early chapters led me to expect that the book would represent the Stoics as providing an interesting set of such arguments designed to help us avoid assent to judgments constituting the emotions we do not want. But what we actually get along these lines turns out to be very little.

Some of the ideas referred to never get detailed discussion at any point in the book, and as far as I can imagine would work by themselves in only a very limited range of cases. Thus the proposal to ask yourself when you are inclined to distress over something, 'Is it bad in itself or merely unexpected?' is not going to do anything even for someone who has broken his leg—an ordinary person presumably thinks this *is* bad in itself. Full-strength Stoicism, with its claim that only virtue matters so that the mere breaking of a leg is indifferent, addresses this point; but that is the very medicine Sorabji wants to reject.

Another strategy to which Sorabji repeatedly refers is to consider, when inclined to become angry, how ugly anger makes you. He himself says once: "I am not sure how often this works" (139). Indeed, the strategy is useful in dealing with minor annoyances such as having to wait longer than agreed on a street corner to meet a friend. But it seems powerless in the face of any injury supposed to be significant: When inclined to outrage over the murder or even mugging of a sister, who would be swayed by reflecting that anger makes one uglier? Anyone open to that (Sir Walter Elliott in *Persuasion*?) would have a value system Sorabji can hardly recommend, and which would bring with it inevitable distress over challenges to vanity.

In fact, it turns out that much of the therapeutic material is not peculiarly Stoic, and so also not specially tied to the Chrysippan analysis of emotion in terms of judgment. One therapeutic exercise that gets somewhat longer treatment (214-16) and is Stoic comes from Epictetus and his teacher Musonius Rufus. But Epictetus's and Musonius's urging us to care for nothing that is not up to us is based on an indifference doctrine, and so is foreign to Sorabji's proprietary dialectical project. As a variant on this strategy of trying to use bits of Stoicism without the foundational parts, I would like to suggest that it can be helpful to read the full-strength doctrines of an Epictetus even if one is not prepared to accept them dogmatically: one rests while reading in the Stoic frame of mind, according to which none of the things one is inclined to get in a fuss over (recognition, beauty, wealth...) matters.

What of the bearing on Sorabji's program of his view that Chrysippus's analysis is not actually correct? The book in fact contains little exploration of any merits of the analysis (beyond a gesture at Martha Nussbaum as defending it powerfully (93, footnote 1)), but gives a lot of play to attacks on it. We have the ancient Posidonius, a dissenting Stoic, who developed examples designed to show that the judgments Chrysippus identifies are neither necessary nor sufficient for emotion; Sorabji's own examples in this spirit; and the modern brain researchers LeDoux and Damasio, whose investigations into the role of the amygdala in fear Sorabji takes to show why it is so easy to come up with the counterexamples (as well as why, in some cases, changing cognitive patterns can reduce unwanted emotion). Yet the dispersal of these discussions throughout the book prevents some desired moments of intertextuality.

Thus Sorabji, in his initial set of objections to the Stoics, contributes the following example to indicate that making the Chrysippan judgments is insufficient for emotion: a person running from a fire might judge that something bad and to be fled from was at hand, and yet only be overcome by the feeling of fear once in a position of safety (115, 137). I wished for a return to discussion of this point in light of the mention later of a distinction from LeDoux and Damasio between being in an emotional state and feeling the emotion (147-48). Could one use that distinction to insist that the person in the original example is in a state of fear all right, but just does not *feel* it? Possibly, Sorabji assumes that feeling cannot come apart from being in the state when the state is a matter of cognition in the way it can when the state is a set of autonomic responses, but it would have been helpful if he had spelled this out.

In general, the book concentrates on giving and defending readings of the texts it considers. There is less room to develop discussion of the views so presented. Questions of the role of introspection in theorizing about emotion, of whether and how a reductionist theory can work, and of how a theory should connect its candidate for what emotion turns out to be with the original phenomenology get no direct treatment. Thus the book's real value may well be in its collecting mission rather than in any sustained argument of its own. We may see the handsome cover photograph (showing, on the intended interpretation, the philosopher Crates demonstrating his contempt for worldly goods by throwing away a basketful of precious jewels) as depicting Sorabji who, having assembled a treasure chest of insights on the emotions, shakes them out for us to take up or not, as we choose.

CONSTANCE C. MEINWALD

University of Illinois/Chicago

THE JOURNAL OF
PHILOSOPHY

The Journal is pleased to announce to our readers and subscribers the launch of its new website, which will have the pages listed below. The address is: *www.journalofphilosophy.org* or simply *www.jphil.org*.

~•~

Purpose & History
Contacts and Addresses
Board of Trustees
Editorial Board
Memos to Authors
Subscriptions & Advertising
Recent & Forthcoming Issues
Electronic Issues
Index 1904–present
Rights and Permissions
Lectures & Conferences
Philosophy Links

~•~

Critical Inquiry

Editor W. J. T. Mitchell
Executive Editor Arnold I. Davidson

A theory driven journal ...

A challenging look ...
Founded 27 years ago, *Critical Inquiry* is an interdisciplinary journal devoted to publishing the best critical thought in the arts and humanities. Combining a commitment to rigorous scholarship with a vital concern for dialogue and debate, the journal presents articles by eminent critics, scholars, and artists on a wide variety of issues central to contemporary criticism and culture.

At scholarly criticism and culture
In *CI* new ideas and reconsideration of those traditional in criticism and culture are granted voice. International authors express inspiring and often provocative ideas on hegemony and variant cultural styles in literature, art, architecture, film, history, music, philosophy, science, and more. Counterparts debate definitions and reflections of the most brilliant, infamous, and popular moments, past and present. In *CI*, authors entertain and challenge while expounding on such issues as Einstein's philosophy of time, early modern women's writing, political-religious identity in the Middle East, and other rigorous subjects. *CI* comes full circle with the electrically charged debates between contributors and their critics.

Order securely through the *CI* Web site
http://www.journals.uchicago.edu/CI/

Critical Inquiry (ISSN 0093-1896) is published quarterly. Regular one-year subscription rates are $37 individuals, $25 students, $105 institutions. Single copies: $8 individuals, $26.25 institutions. Outside the US, please add $5 postage. Canadian residents please add 7% GST and $5 postage. MasterCard and Visa payments accepted. Credit card customers may fax orders to (773) 753-0811. To order your subscription send check, purchase order, or credit card type, account #, exp. date, telephone #, and signature to the University of Chicago Press, Journals Division, Department SS1SA, P.O. Box 37005, Chicago, IL 60637 USA. For more information, please visit our Web site at *http://www.journals.uchicago.edu/CI/.*

The University of Chicago Press • Journals Division • Box 37005 • Chicago, IL 60637

March 2001

New Philosophy Titles from . . .

Prometheus Books

SKEPTICAL PHILOSOPHY FOR EVERYONE
Richard H. Popkin and Avrum Stroll
325 pp • ISBN 1-57392-936-0 • HC $35

VIRTUE ETHICS
An Introduction
Richard Taylor
141 pp • ISBN 1-57392-943-3 • PB $18

PORNOGRAPHY, SEX, AND FEMINISM
Alan Soble
210 pp • ISBN 1-57392-944-1 • HC $28

. . . and

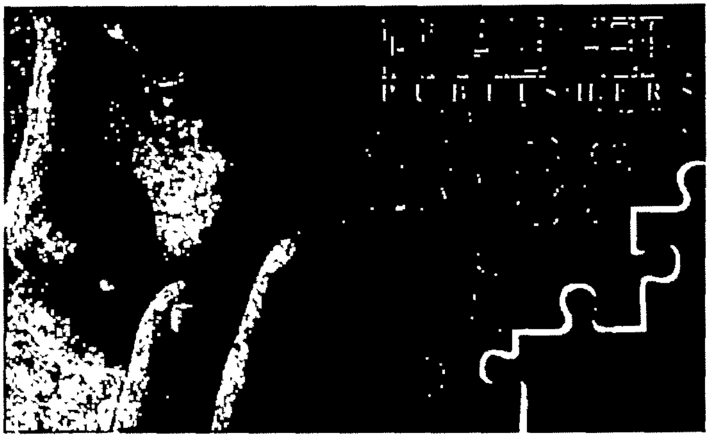
Humanity Books

PATRIOTISM
Edited by Igor Primoratz
276 pp • ISBN 1-57392-955-7 • PB $29

NEW ESSAYS IN FICHTE'S *FOUNDATION OF THE ENTIRE DOCTRINE OF SCIENTIFIC KNOWLEDGE*
Edited by Daniel Breazeale and Tom Rockmore
277 pp • ISBN 1-57392-914-X • HC $55

REFRAMING THE FRAME OF REASON
"Trans-lation" *in* and *beyond* Kant and Derrida
Ulrike Oudée Dünkelsbühler
Translated by Max Statkiewicz, Foreword by Jacques Derrida
275 pp • ISBN 1-57392-958-1 • HC $65

For exam copies contact editorial@prometheusbooks.com
Call toll free [illegible] • Fax [illegible] • www.prometheusbooks.com



The Death and Resurrection of the Author?

Edited by William Irwin

Contributions in Philosophy, No 83

"Provocative contributions by both analytic and continental thinkers to two of the central aspects of current debate—namely, the death of the author and the nature and function of author constructs."

Jeff Mitscherling,
University of Guelph, Ontario

Greenwood Press • 2002
0-313-31870-0 • $64.95

A paperback edition is available:
0-313-32275-9 • $23.95

Greenwood
PUBLISHING GROUP

Praeger and Greenwood Press are Imprints of the Greenwood Publishing Group

88 Post Road West,
P.O. Box 5007
Westport, CT 06881-5007
Order Fax (203) 750-9790

Paradigms Explained

Rethinking Thomas Kuhn's Philosophy of Science

By Erich von Dietze

"The best analytical study of Thomas Kuhn's paradigm theory that has been written . . . a seminal and original contribution to the growing world-wide interest in Kuhn studies."

William A. Johnson,
Brandeis University

Praeger Publishers • 2001 • 192 pages
0-275-96999-1 • $64.00

Refuting Peter Singer's Ethical Theory

The Importance of Human Dignity

By Susan Lufkin Krantz

Shows how Singer's ethical theories threaten human values in a variety of ways. Topics include abortion and infanticide, euthanasia, and assisted suicide.

Praeger Publishers • 2002 • 152 pages
0-275-97083-3 • $54.95

PLACE YOUR ORDER TOLL-FREE: 1-800-225-5800
www.greenwood.com

ROUTLEDGE

NEW IN PHILOSOPHY

ACTS OF RELIGION
Jacques Derrida
Edited with an introduction by Gil Anidjar

"This important collection of essays traces the development of Derrida's long-standing interest in religion into a virtual obsession. In ways difficult to calculate Derrida has become one of the most provocative 'religious' thinkers of our time."
— Mark C. Taylor, author of *The Moment of Complexity, Emerging Network Culture and Grave Matters*

448 pp • $24.95

ARGUING ABOUT ART
Contemporary Philosophical Debates
Second Edition
Edited by Alex Neill and Aaron Ridley

This is an expanded and revised new edition of a highly acclaimed anthology. It presents twenty-seven readings in a clear and accessible format, discussing the major themes and arguments in aesthetics.

496 pp • $27.95

CRITICAL THINKING
A Concise Guide
Tracy Bowell and Gary Kemp

This much-needed guide to thinking skills provides a thorough introduction to thinking clearly and rationally for oneself. Accessibly written, this book equips readers with the essential skills required to tell a good argument from a bad one.

288 pp • $20.95

METAPHYSICS
A Contemporary Introduction
Second Edition
Michael J. Loux

In this fully revised and updated version of the highly successful first edition, Michael J. Loux provides a fresh look at the central topics in metaphysics, rendering this essential reading for anyone interested in metaphysics.

320 pp • $22.95

ROUTLEDGE PHILOSOPHY GUIDEBOOK TO HEGEL AND *THE PHENOMENOLOGY OF SPIRIT*
Robert Stern

"[Stern's] exposition is clear, concise and to the point. It cuts through the thicket of scholarly dispute and provides a shrewd and commonsensical reading of the text. This is the best introductory guide to the Phenomenology in English."
—Fred Beiser, University of Indiana

256 pp • $14.95

THEORIES OF DEMOCRACY
A Critical Introduction
Frank Cunningham

"A very full and descriptive treatment of theories about democracy. It embraces a very wide range of position and supports each position by a great mass of reference and reading, making it, I think, a unique guide to the subject."
—Ross Harrison, University of Cambridge

256 pp • $21.95

www.routledge-ny.com

1.800.634.7064

ROUTLEDGE · Taylor & Francis Group

Best in Philosophy

Cicero: *On Moral Ends*
Marcus Tullius Cicero
Julia Annas, Editor
Cambridge Texts in the History of Philosophy
0-521-66061-0, Hardback, $38 00
0-521-66901-4, Paperback, $14.00

Condillac: *Essay on the Origin of Human Knowledge*
Etienne Bonnot de Condillac
Hans Aarsleff, Editor
Cambridge Texts in the History of Philosophy
0-521-58467-1, Hardback, $60.00
0-521-58576-7, Paperback, $22 00

Intellectual Trust in Oneself and Others
Richard Foley
Cambridge Studies in Philosophy
0-521-79308-4, Hardback, $55.00

Physicalism and Its Discontents
Carl Gillett and
Barry Loewer, Editors
0-521-80175-3, Hardback, $60 00

Practical Rules
When We Need Them and When We Don't
Alan H. Goldman
Cambridge Studies in Philosophy
0-521-80729-8, Hardback, $55.00

The Structure of Values and Norms
Sven Ove Hansson
Cambridge Studies in Probability, Induction and Decision Theory
0-521-79204-5, Hardback, $60 00

Divine Hiddenness
New Essays
Daniel Howard-Snyder and
Paul K. Moser, Editors
0-521-80353-5, Hardback, $60.00
0-521-00610-4, Paperback, $22.00

G.E. Moore's Ethical Theory
Resistance and Reconciliation
Brian Hutchinson
0-521-80055-2, Hardback, $55 00

Wittgenstein
Biography and Philosophy
James C. Klagge, Editor
0-521-80397-7, Hardback, $55.00
0-521-00868-9, Paperback, $20.00

Fichte
The Self and the Calling of Philosophy, 1762-1799
Anthony J. La Vopa
0-521-79145-6, Hardback, $55.00

Enlightenment and Action from Descartes to Kant
Passionate Thought
Michael Losonsky
0-521-80612-7, Hardback, $55.00

Evil and the Augustinian Tradition
Charles T. Mathewes
0-521-80715-8, Hardback, $60.00

Collective Rationality and Collective Reasoning
Christopher McMahon
0-521-80462-0, Hardback, $55.00
0-521-01178-7, Paperback, $20.00

Social Action
A Teleological Account
Seumas Miller
0-521-78316-X, Hardback, $55.00
0-521-78886-2, Hardback, $20.00

Practical Rationality and Preference
Essays for David Gauthier
Christopher W. Morris and
Arthur Ripstein, Editors
0-521-78184-1, Hardback, $55.00

Best in Philosophy from Cambridge

Contemporary Philosophy in Focus

Announcing
a new series

Contemporary Philosophy in Focus is a series of introductory volumes to many of the dominant philosophical thinkers of the current age. Each volume consists of newly commissioned essays that cover all the major contributions of a preeminent philosopher in a systematic and accessible manner.

Daniel Dennett
Andrew Brook and
Don Ross, Editors
0-521-80394-2 Hardback $60.00
0-521-00864-6 Paperback $20.00

Robert Nozick
David Schmidtz, Editor
0-521-78226-0 Hardback $60.00
0-521-00671-6 Paperback $20.00

Practical Reasoning in a Social World
How We Act Together
Keith Graham
0-521-80378-0 Hardback $55.00

Nietzsche: *Beyond Good and Evil*
Friedrich Nietzsche
Introduction by
Rolf-Peter Horstmann
Translated by Judith Norman
Cambridge Texts in the History of Philosophy
0-521-77078-5 Hardback $38.00
0-521-77913-8 Paperback $14.00

Wittgenstein's *Tractatus*
A Dialectical Interpretation
Matthew B. Ostrow
0-521-80936-3 Hardback $55.00
0-521-00649-X Paperback $20.00

The Nature of Consciousness
Mark Rowlands
0-521-80858-8 Hardback $60.00

Real Ethics
Reconsidering the Foundations of Morality
John M. Rist
0-521-80921-5 Hardback $65.00
0-521-00608-2 Paperback $23.00

Adam Smith: *The Theory of Moral Sentiments*
Knud Haakonssen, Editor
Cambridge Texts in the History of Philosophy
0-521-59150-3 Hardback $65.00
0-521-59847-8 Paperback $23.00

Now in paperback...

From Chance to Choice
Genetics and Justice
Allen Buchanan, Dan Brock,
Norman Daniels, and
Daniel Wikler
0-521-66977-4 Paperback $23.00

CAMBRIDGE
UNIVERSITY PRESS
[illegible]

✓

THE JOURNAL OF PHILOSOPHY

18/7/02

VOLUME XCIX, NUMBER 4
APRIL 2002

page
167 *Egalitarianism against the Veil of Ignorance* John E. Roemer
185 *Induction, Hume, and Probability* Dugald Murdoch

COMMENTS AND CRITICISM
200 *Amazing Knowledge* Stephen Schiffer

BOOK REVIEWS
203 Deborah K. W. Modrak: *Aristotle's Theory of Language and Meaning* Richard Tierney

210 NEW BOOKS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2002

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 2002 by the Journal of Philosophy, Inc. ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY
VOLUME XCIX, NO. 4, APRIL 2002

EGALITARIANISM AGAINST THE VEIL OF IGNORANCE*

The construct of a veil of ignorance has been used as a powerful tool of analysis in modern theories of distributive justice, in various forms, most prominently by three authors: John Harsanyi, John Rawls, and Ronald Dworkin.[1] Of these, Rawls has claimed that the veil-of-ignorance argumentation leads to a kind of egalitarianism; Harsanyi, while proclaiming himself a utilitarian, believed that the kind of utilitarianism produced by his veil of ignorance would not be offensively inegalitarian; and Dworkin, although not using his veil of ignorance to justify resource egalitarianism, does use it to deduce or compute what, he claims, the resource-egalitarian distribution *is*.

My aim, here, is to argue that veil-of-ignorance arguments, when properly done, contravene fundamental egalitarian principles. If I am right, and if the veil of ignorance is the right tool for deducing what distributive justice requires, then those of us who advocate egalitarianism must cease to do so. It will perhaps not surprise the reader to learn that I shall, instead, indicate in the conclusion why I believe the veil of ignorance is the wrong tool for deducing the requirements of distributive justice. That final argument is, however, more tentative in nature than what precedes it.

I have already written elsewhere[2] that I think Harsanyi's *impartial observer* (IO) argument, in which he deduces how a representative

---

* I am grateful to G. A. Cohen, Mathias Risse, and Joaquim Silvestre for their comments on an earlier draft.

[1] Harsanyi, "Cardinal Utility in Welfare Economics and the Theory of Risk-taking," *Journal of Political Economy*, LXI (1953): 434-35; Rawls, *A Theory of Justice* (Cambridge: Harvard, 1971); Dworkin, "What Is Equality? Part 2: Equality of Resources," *Philosophy and Public Affairs*, X (1981): 283-345.

[2] *Theories of Distributive Justice* (Cambridge: Harvard, 1996), chapter 4.

0022-362X/02/9904/167–84

soul, behind an appropriate veil of ignorance, should solve the decision problem of allocating wealth among individuals in the world, is basically correct: what I mean is that, given the premise that the veil of ignorance is the right tool for the job, then Harsanyi has done the decision theory properly. Two comments are in order. First, Harsanyi put a false gloss on his conclusion—namely, that the IO should be described as "utilitarian." This resulted from Harsanyi's confusion between the maximization of expected utility of one decision maker, and the maximization of the average utility of a set of individuals.[3] The former makes sense with an (ordinal) preference relation over lotteries, and the latter makes sense only when utility is interpersonally comparable to a degree—it must be cardinally measurable and unit comparable. My claim is that Harsanyi's decision theory is correct, but he is wrong to call his soul a utilitarian: indeed, utilitarianism is not a coherent concept, given the information postulated by Harsanyi. It was unsurprising that he erred in this way, as clarity on the problem of the measurability and comparability of utility was not achieved until the 1970s, two decades after Harsanyi wrote.[4]

Second, Harsanyi does not, in fact, completely solve the decision problem of the IO behind the veil of ignorance; rather, he deduces that the IO's optimization problem must be one of a continuum of possible optimization problems—that is, that it must maximize *some* positive linear combination of the von Neumann-Morgenstern utility functions of the individuals in question. But Harsanyi provides no (correct) argument for what the choice of those weights should be. That choice is indeterminate in his environment.

I, and many others, have written that Rawls's difference principle does not flow from any conventional decision problem of a soul in his original position. The issue here is whether it is appropriate to take such a soul as being infinitely risk averse, for that seems to be the only sensible way of deducing the difference principle. But, apart from his exotic decision theory, Rawls's original position seems to be incor-

[3] The point was made initially by Amartya Sen, "Non-linear Social Welfare Functions: A Reply to Professor Harsanyi," in Robert E. Butts and Jaakko Hintikka, eds., *Foundational Problems in the Social Sciences* (Boston: Reidel, 1977), pp. 297-301; and later elaborated upon by John Weymark, "A Reconsideration of the Harsanyi-Sen Debate on Utilitarianism," in Jon Elster and Roemer, eds., *Interpersonal Comparisons of Well-being* (New York: Cambridge, 1991), pp. 255-320.

[4] These later developments are surveyed in my *Theories of Distributive Justice*, chapter 1.

rectly specified for *his* purposes—at least I have argued this point.[5] The claim is that, in the original position, morally arbitrary aspects of persons' characteristics (such as the wealth they are assigned through the families they inherit in the birth lottery) are treated identically to non-morally-arbitrary characteristics (principally, their plans of life), and such a move cannot reflect Rawls's views. I shall not repeat that argument here.

Dworkin (*op. cit.*) constructs a "thin" veil of ignorance, where the souls behind the veil know the preferences of those individuals whom they represent, but do not know those individuals' resource endowments. Unlike Harsanyi and Rawls, the activity behind the Dworkinian veil is not a decision problem of an individual soul, but is a market among many souls.[6] I have argued[7] that, although Dworkin's own conception of how this insurance market should be modeled is wrong, it is a straightforward exercise in general equilibrium theory to model properly the insurance market behind Dworkin's veil. I shall review that exercise in section III below. It turns out that, when properly modeled, the (modified) Dworkinian insurance market can produce—and will in general produce—some unpleasant results, which, I claim, disqualify it as a mechanism for implementing "equality of resources." Indeed, we shall see that the problem with Dworkin's mechanism is related to the problem diagnosed in the generalization of Harsanyi's mechanism.

In the next section, I shall present what I think is a compelling solution to a representative soul's decision problem, in a particularly simple environment. I shall build upon Harsanyi. To be specific, I shall append one new postulate to Harsanyi's IO argument, and thereby deduce the *unique* decision problem that the soul should solve. I shall then study, in a series of examples, the distribution of wealth that follows from this solution, and argue that it contravenes an important egalitarian postulate. Thus, I claim that a properly modeled decision problem behind the veil of ignorance cannot be used to justify egalitarianism.

Some might say this is no surprise—they already knew. I disagree. Despite the qualifications that their authors have added over the years, it is, I think, unquestionable that serious thinkers with an egalitarian bias have often hoped to justify their ethos with veil-of-ignorance arguments. In my view, this includes Dworkin, who, although claiming that his veil is used only to compute what egalitarianism requires and not to argue for it, nevertheless hopes that the hypothetical insurance market he constructs behind a thin veil of ignorance will render his conception of resource egalitarianism appealing. My intention is to show that such hopes are fruitless.

---

[5] See my *Theories of Distributive Justice*, chapter 5, and my "Three Egalitarian Views and American Law," *Law and Philosophy*, XX (2001): 433-60.

[6] Although Rawls claims the activity behind his veil is a "social contract," in fact, formally speaking, it is a decision problem of a single soul, as all souls in the Rawlsian original position are identical.

[7] "Equality of Talent," *Economics and Philosophy*, I (1985): 151-81; *Theories of Distributive Justice*, chapter 7; and "Three Egalitarian Views and American Law."

I. HARSANYI REFINED

There is a population of individuals, defined as a set of types, identified by a parameter $\omega \in \Omega$, where Ω is a sample space; types are distributed according to a probability measure F on Ω. Each type is characterized by two functions of wealth, denoted $u(\cdot, \omega)$ and $v(\cdot, \omega)$. $u(\cdot, \omega)$ is a von Neumann-Morgenstern (vNM) utility function representing type ω's preferences over wealth lotteries, and $\{v(\cdot, \omega) | \omega \in \Omega\}$ is a profile of functions defined on wealth which measures the welfare of individuals in an interpersonally comparable manner. Thus, $v(W^1, \omega^1) = v(W^2, \omega^2)$ means that a type ω^1 individual with wealth level W^1 and a type ω^2 individual with wealth level W^2 enjoy the same welfare level.

The only consistency requirement, with regard to the functions u and v is that, for each ω, they give the same orderings over wealth levels. Thus, one possible lottery is a sure thing, where an individual receives a given level of wealth with probability one, and the von Neumann-Morgenstern preferences of the individual must rank these sure-thing lotteries according to the individual's *welfare*. That is to say:

$$(\forall \omega \in \Omega)(\forall W, W')(u(W, \omega) \geq u(W', \omega) \Leftrightarrow v(W, \omega) \geq v(W', \omega))$$

If we assume that u and v are strictly monotone increasing in W, then this requirement is trivially satisfied. (Of course, it would not be a trivial statement, were wealth a multidimensional vector.)

In the language of social-choice theory, we are given, for each individual, his ordinal preferences on wealth lotteries, which are assumed to obey the vNM axioms, and an interpersonally level-comparable measure of welfare. It is noteworthy that we lack sufficient information, in this environment, for utilitarianism to be coherent, for we have no *unit comparable* measure of welfare. (Of course, one could endow the profile v with unit comparability as well, and then utilitarianism would be coherent. But such information is unnecessary for the argument to follow.)

We shall construct the vNM utility function of the soul that represents the generic member of Ω, that is, the soul whose job it will be to decide upon the distribution of wealth in the world, knowing only that it could become, in the birth lottery, any of the actual types in the population, according to the probability distribution F. Let a *prospect* be an ordered pair (W, ω). We assume that the soul has a preference order over lotteries of prospects, which obeys the vNM axioms. Let U be the soul's vNM utility function on prospects, that is the vNM utility function that represents the soul's preferences over such lotteries. We assume, following Harsanyi (*op. cit.*):

> (A1) *Principle of acceptance:* on the domain of wealth lotteries over prospects involving only a single type ω, $u(W, \omega)$ represents the soul's vNM preferences.

(A1) is Harsanyi's principle of acceptance. It is justified by saying that, if the soul were asked to choose between two wealth lotteries in which only the wealth of a single individual, of some given type, is at issue, then it should choose the lottery that the individual in question would choose. This, at least, is required for the soul to be "representative."

It thus follows from the vNM representation theorem that there are functions $\alpha(\omega)$ and $\beta(\omega)$, with $\alpha(\omega) > 0$, such that:

$$(2.1) \quad \forall\, \omega \in \Omega \quad U(W, \omega) = \alpha(\omega)u(W, \omega) + \beta(\omega)$$

How does the soul choose between prospects involving the possibility of its becoming persons of two different types? We assert:

> (A2) *Principle of neutrality:* let L be a lottery with over two prospects, (W, ω) and (W', ω'). The soul is indifferent between the two prospects if the types in those prospects enjoy equal welfare levels at the wealths in question; that is, if $v(W, \omega) = v(W', \omega')$.

This principle is one of neutrality, for it says that the soul imposes no external preferences over the nature of the lives led by individuals of different types: it is concerned only with the universal concept of well-being called welfare. Were the soul to base its choice over prospects involving different types on a view that some kinds of life were better than other kinds, for reasons not related to the welfare experienced by those types, it would, I think, be shirking its duty to represent all types. For by saying that two individuals enjoy the same level of welfare, I mean to say that they rate their lives as being equally good.

Now let $W^1(\omega)$ be an allocation of wealth to types at which all types enjoy the same welfare, that is:

$$\forall\, \omega \in \Omega \quad v(W^1(\omega), \omega) = k^1$$

and let $W^2(\omega)$ be another wealth allocation at which all types enjoy the same level of welfare, which is greater than they enjoy at the first allocation, that is:

$$\forall\, \omega \in \Omega \quad v(W^2(\omega), \omega) = k^2 > k^1$$

(These two distributions of wealth need not be feasible for the society in question.) By (A2), we have:

$$\begin{aligned} (2.2) \quad \forall\, \omega \in \Omega \quad & \alpha(\omega)u(W^1(\omega), \omega) + \beta(\omega) = k^{1\prime} \\ & \alpha(\omega)u(W^2(\omega), \omega) + \beta(\omega) = k^{2\prime} \end{aligned}$$

for some constants $k^{2\prime} > k^{1\prime}$.

Now we are free to choose any vNM utility function from a positive affine family to represent type ω's vNM preferences over wealth lotteries; we now suppose that we have chosen the profile $u(\,\cdot\,, \omega)$ which renders $k^{1\prime} = 0$ and $k^{2\prime} = 1$. (This, indeed, fixes the function u.) Then we can solve (2.2) for the functions α and β:

$$(2.3) \quad \alpha(\omega) = \frac{1}{D(\omega)}, \quad \beta(\omega) = \frac{-u(W^1(\omega), \omega)}{D(\omega)}$$

where

$$D(\omega) = u(W^2(\omega), \omega) - u(W^1(\omega), \omega)$$

and so we have:

$$(2.4) \quad U(W, \omega) = \frac{u(W, \omega)}{D(\omega)} - \frac{u(W^1(\omega), \omega)}{D(\omega)}$$

We have thus completely determined the souls' vNM preferences on lotteries over prospects, for by the vNM axioms, it suffices to compute its utility function on the pure prospects themselves, which we display in (2.4). It is worth observing that the vNM preferences represented by U are independent of the particular profile u we chose to represent the preferences of individual types. If we replace each function $u(\,\cdot\,, \omega)$ with a positive affine transfer of itself, the function U is unchanged (note that the various constants factor out of both numerator and denominator in (2.4)). Thus, the function U depends only on the vNM *preferences* of the types, on the function v, and on the two wealth distributions chosen.

But this raises a final issue. The determination of U apparently depends upon the particular wealth distributions W^1 and W^2 chosen. It is, however, easy to verify that, if we chose two other wealth distributions at which welfare is equal across individuals, then the vNM utility function derived for the soul is equivalent to U. Thus, the preferences over prospects represented by U are in fact well defined by axioms (A1) and (A2).

We now analyze the soul's decision problem, which is to choose a distribution of wealth among types. The soul chooses that feasible distribution of wealth $W(\omega)$ that maximizes its expected utility when facing the 'birth lottery' F. We shall assume that average wealth in the society is given at $\overline{W}$, and is independent of how wealth is distributed. (This is not necessary, but will simplify our examples.) Then the soul's problem is to choose a distribution of wealth $W(\cdot)$ to:

$$(2.5) \quad \underset{W(\cdot)}{\text{Max}} \int U(W(\omega), \omega) dF(\omega)$$

$$s.t. \int W(\omega) dF(\omega) = \overline{W}$$

If u is concave in W, then U is concave in W, and the solution to program (2.5) entails:

$$\text{(C1) for some } \lambda > 0, \frac{u'(W(\omega), \omega)}{D(\omega)} \leq \lambda$$

with equality holding when $W(\omega) > 0$ and

$$\text{(C2)} \int W(\omega) dF(\omega) = \overline{W}$$

where u' denotes the derivative of u with respect to W.

We thus assert that we have solved the problem of deciding upon the distribution of wealth using the veil-of-ignorance approach—at least in the simple environment postulated here (where all types are risk neutral or risk averse [thus concave vNM utilities], average wealth is independent of wealth's distribution, only welfare matters, and so on).

II. SOME EXAMPLES

We next compute the soul's choice of wealth distribution for some examples.

Example 1. All types are risk neutral.

In this case, because we are at liberty to choose any vNM utility function from among a type's positive affine family, we choose:

$$\forall \omega \in \Omega \quad u(W, \omega) = W$$

Then (C1) becomes

$$\frac{1}{W^2(\omega) - W^1(\omega)} \leq \lambda, \text{ with equality when } W^*(\omega) > 0$$

(We denote by $W^*(\omega)$ the optimal solution to (2.5).) It follows that

$$\lambda = \operatorname*{Sup}_{\omega} \frac{1}{W^2(\omega) - W^1(\omega)}$$

and so

$$W^*(\omega) = 0, \text{ except when } \omega = \operatorname*{Arg\,Inf}_{\omega} (W^2(\omega) - W^1(\omega))$$

Thus, the only types who receive any wealth at the optimal distribution are those for whom the wealth difference $(W^2(\omega) - W^1(\omega))$ is minimal.

It is reasonable to assert, I think, that:

(A3a) If ω is disabled and ω' is able, but the two types are in other ways similar, then $W^2(\omega) - W^1(\omega) > W^2(\omega') - W^1(\omega')$

That is, to effect a welfare saltus between two given welfare levels, for two individuals who are similar except for a disability, a larger wealth increment is needed for the disabled person.

Indeed, the posture that we implicitly adopt in (A3a) is that we know what disability means, and (A3a) asserts a relationship between disability and the production of welfare. But we could stand differently, and take (A3a) as the *definition* of disability.

In like manner, we might assert:

(A3b) If ω has expensive tastes and ω' has cheap tastes, but the two types are in other ways similar, then

$$W^2(\omega) - W^1(\omega) > W^2(\omega') - W^1(\omega')$$

If we accept (A3a) and (A3b), either as reasonable properties of disability and expensive tastes, or as definitions of those properties, then, in example 1, the only types who receive any wealth at the optimal distribution are those who are able and/or have cheap tastes.

Example 2. There is a fixed number p, $0 < p < 1$, such that

$$\forall \omega \in \Omega \quad u(W, \omega) = \frac{W^{1-p}}{1-p}$$

This is the constant-relative-risk-aversion utility function. All types are equally risk averse in this example. We take

$\Omega = [0, 1]$, $W^2(\omega) = k^2\omega$, $W^1(\omega) = k^1\omega$, $k^2 > k^1$

Assuming, for the moment, an interior solution where everyone receives positive wealth, (C1) reduces to

$$\frac{W^*(\omega)^{-p}(1-p)}{D_1(\omega)} = \lambda$$

where $D_1(\omega) = W^2(\omega)^{1-p} - W^1(\omega)^{1-p}$, and so

$$W^*(\omega) = \left(\frac{1-p}{\lambda}\right)^{1/p} D_1(\omega)^{-(1/p)} \tag{2.6}$$

λ may now be chosen to render (C2) true, and so (2.6) is indeed the optimal solution. Note that $D_1(\omega)^{-(1/p)}$ is monotone decreasing in $D_1(\omega)$. Since $W^2(\omega) - W^1(\omega) = (k^2 - k^1)\omega$ is increasing in ω, we identify larger ω with more disability. But $D_1(\omega) = \omega^{1-p}(k^{2^{1-p}} - k^{1^{1-p}})$ is also increasing in ω. Thus we have, in this example, that the able receive more wealth at the chosen distribution than the disabled.

Example 3. $\Omega = [0, 1]$, F is the uniform distribution, and

$$u(W, \omega) = \frac{W^{1-\omega}}{1-\omega}$$

In this example, types have different but constant relative degrees of risk aversion, *and* a person's type is completely specified by his risk preferences. (That is, all individuals with the same risk preferences have the same welfare function $v(\,\cdot\,, \omega)$.)

To enable us to compute a solution in this example, we need to specialize further. We show:

Proposition. In example 3, *if*

$$W^2(\omega) = \omega^{(2\omega)/(1-\omega)} \text{ and } W^1(\omega) = \omega^{(2\omega+1)/(1-\omega)} \tag{2.7}$$

then

$$W^*(\omega) = \frac{1}{\omega^2}\left(\frac{1}{\lambda}\right)^{1/\omega}$$

where λ *is the solution of the equation* $\lambda \operatorname{Log} \lambda = \overline{W}^{-1}$

Proof:

1. (C1) implies

$$W^*(\omega) = \left(\frac{1-\omega}{\lambda D_2(\omega)}\right)^{1/\omega}$$

where $D_2(\omega) = W^2(\omega)^{1-\omega} - W^1(\omega)^{1-\omega}$

Using (2.7) we have:

$$D_\gamma(\omega) = (1 - \omega)\omega^{2\omega}$$

and we hence compute

$$W^*(\omega) = \lambda^{-(1/\omega)} \frac{1}{\omega^2}$$

Now (C2) in this example becomes:

$$(2.8) \quad \int_0^1 \lambda^{-(1/\omega)} \frac{1}{\omega^2}\, d\omega = \overline{W}$$

Let $q = -(1/\omega)$ be a change of variable; since $dq/d\omega = 1/\omega^2$, we may transform (2.8) to:

$$\int_{-\infty}^{-1} \lambda^q dq = \overline{W}$$

Integrating gives:

$$(2.9) \quad \left.\frac{\lambda^q}{\mathrm{Log}\, \lambda}\right|_{-\infty}^{-1} = \overline{W}$$

if $\lambda > 1$, (2.9) evaluates to:

$$\frac{1}{\lambda\, \mathrm{Log}\, \lambda} = \overline{W}$$

as claimed in the proposition's statement. Now $\overline{W} > 0$ tells us that indeed $\lambda > 1$, and the proposition is proved.

In figure 1, we plot the functions W^2, W^1, and W^* of the Proposition in the (ω, wealth) plane. The higher and lower thin lines are the graphs of the functions W^2 and W^1, respectively. Each heavy line is the solution W^* for a particular value of $\overline{W}$, with higher average wealths associated with the higher graphs. We see that the low ω types require a large wealth increment to make the saltus between the two welfare levels, while the high ω types require a much smaller wealth increment. We therefore identify low ωs with the disabled. The optimal distribution is nonmonotonic, but it is generally true at all levels of average wealth that the most able receive greater wealth than the most disabled. The biggest wealths, however, go to types in the mid-range of "ability."

In this example, low ω is associated both with disability, and with a low degree of risk aversion ($\omega = 0$ is risk neutral). Thus, with an equal wealth distribution as benchmark, wealth is redistributed from dis-

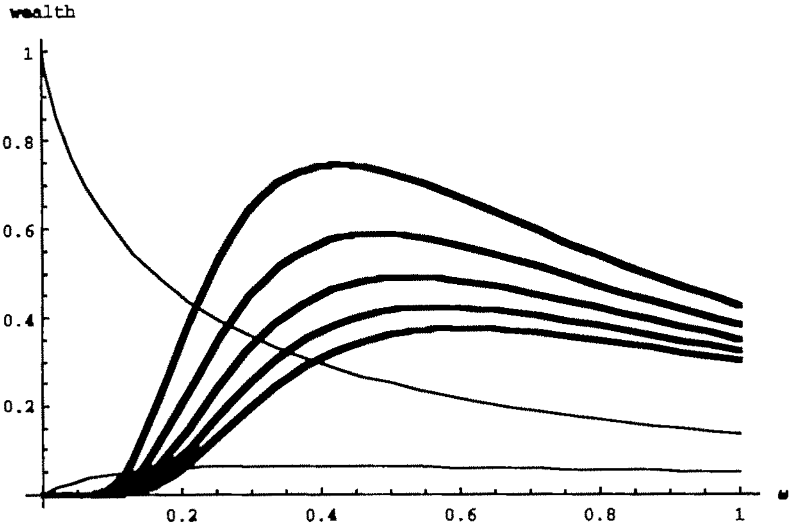


Figure 1: Graphs of the functions of the Proposition

abled types who are almost risk neutral to able types who are highly risk averse.

We cannot say precisely what utilitarianism would recommend in any of these examples, because, as I wrote, we do not have sufficient information to define utilitarianism. Utilitarianism would distribute wealth to maximize the average value of v, but that average value is not well defined without unit comparability of welfare.

III. DWORKIN'S INSURANCE MECHANISM

Dworkin, unlike Rawls or Harsanyi, does not employ a veil of ignorance to argue for his political view, but rather, as he says, to define what it consists in. In brief, Dworkin believes that equality-of-resources is a state in which the transferable resources (principally, money) that individuals have, provide them appropriate compensation for the shortfalls they sustain in their endowments of non-transferable resources. He believes that individuals are responsible for their preferences, but not for their resource endowments, and what distribution of money is "appropriate," in the above sense, must rest upon this fact.

In Dworkin's scheme, individuals are represented by souls behind a veil of ignorance who know the preferences (in particular, the preferences over lotteries) of the persons they represent, but do not know what allocation of resources—internal or external—their persons will receive in the birth lottery. We shall identify external resources with wealth and call internal resources *nontransferable*. Each soul is given the same amount of purchasing power with which to purchase insurance against a bad draw in the birth lottery, which allocates those resources. There will be an equilibrium in this insurance market, which will redistribute money wealth among actual persons, contingent upon the realization of the birth lottery (when insurance contracts pay out). Dworkin recommends a tax scheme in the world that implements the contracts that the souls sign behind his veil. This scheme models Dworkin's view that individuals should be responsible for their preferences (as the souls employ those preferences when choosing contracts), but not for their resources (as all souls have the same purchasing power behind the veil).

There is a canonical way of modeling this scheme using the notion of a market for contingent claims. Imagine that each soul begins with *zero* income behind the veil. (In particular, each has an equal amount of purchasing power.) The probability distribution of resources in the actual world is known to all souls. A *state of the world* is a particular allocation of resources to souls (that is to say, to preference orders). Let s be a state of the world. The commodity X_s is the promise by the insurance company to deliver $1000 to the purchaser of the commodity, should state s occur. If I, a soul, purchase x_s units of X_s, then, if state s occurs, the person I represent receives a check for $\$1000x_s$ from the insurance company. If I sell x_s units of this commodity, then, should state s occur, my person writes a check for $\$1000x_s$ to the insurance company. Feasibility of these contracts requires that all promises to pay can be met with the wealths available to the persons involved in the state in question.

A price vector is a vector $(p_1, \ldots, p_s, \ldots)$, one price for each state, where p_s is the price of one unit of X_s. It is important to note that these commodities are bought and sold before the veil is lifted: the market takes place behind the veil of ignorance, and among souls, not persons. An equilibrium is a price vector and a matrix of demands $\{x_{is}\}$, where x_{is} is the demand by soul i for commodity X_s, at which the total demand for the commodity X_s equals the total supply of this commodity, for every s. (Demands are positive numbers, supplies are negative numbers.) It is important to note that the only way a soul acquires the purchasing power to purchase delivery of dollars in some states is by selling the promise to deliver dollars in other states. Thus, any soul who wishes to buy insurance in some states must sell insurance in other states. Typically, we imagine a soul's selling contracts to deliver dollars in states in which its person is wealthy and well resourced, and purchasing promises to deliver dollars in states in which its person is poor and/or poorly resourced.

Formally, we define an equilibrium as follows. A prospect is an ordered pair (W, s)—the enjoyment of transferable wealth W in state s. In different states, the souls are embodied as persons with different endowments of nontransferable resources and wealth—but a soul always has the preferences of one individual. Let W_{is} be the money wealth of soul i in state s, and let E_{is} be the endowment of nontransferable resources of soul i in state s. Then soul i's vNM utility function is defined on pairs (W, E). Let π_s be the probability that state s occurs, and let the states be enumerated $1, 2, \dots, S$. An equilibrium is a price vector $p = (p_1, \dots, p_S)$ and a matrix of demands $\bar{x} = \{\bar{x}_{is}\}$ such that:

(1) (utility maximization) For each i, the vector of demands $(\bar{x}_{i1}, \dots, \bar{x}_{iS})$ maximizes $\Sigma_s\ \pi_s u_i(W_{is} + x_s, E_{is})$ subject to the budget constraint $\Sigma_s p_s x_s \leq 0$ and the feasibility constraints $W_{is} + x_s \geq 0$, for all s; and
(2) (market clearing) for each s, $\Sigma_i\ x_{is} = 0$

The budget constraint in (1) states that souls must finance the purchase of deliveries of wealth in some states by selling contracts promising to deliver wealth in other states, and the other constraints in (1) say that there is no state in which a soul can promise to deliver more wealth to the insurance company than its person is endowed with, in that state. The post-transfer wealth of the person whom soul i becomes in state s is $W_{is} + \bar{x}_{is}$ (remember that if a person i pays out in state s then $\bar{x}_{is} < 0$).

It is a consequence of the Arrow-Debreu existence theorem that, if the utility functions u_i are quasi-concave (which, informally, we may think of as happening if agents are risk neutral or risk averse), then such an equilibrium exists.[8] It is perhaps important to reiterate that all *economic* activity—that is, the buying and selling of commodities, which are *contingent claims*—happens behind the veil of ignorance, and using the unit of exchange that is recognized there, not real-world dollars. After the veil is lifted, and a state of the world is realized, then deliveries of dollars are made among real persons.

Now, the actual world we live in is one of these states, call it s^*, which, recall, associates preference orders with wealth endowments. At equilibrium, souls hold contracts either to deliver dollars to the insurance company, to receive dollars from the insurance company, or to do neither, should state s^* occur. When these transactions occur, a transfer of dollars among citizens will have transpired. Dworkin's optimal tax system would mimic this transfer.

---

[8] I do not say equilibrium exists only under this condition.

What I have shown[9] is that this tax system will generally behave "pathologically," from a resource-egalitarian viewpoint. The simplest example is the following. There are two persons, Andrea and Bob, and there is one internal resource that we may call "ability." The money wealth endowment of each, in the real world, is $10,000. In the real world, Andrea is able and Bob is disabled—to wit, she possesses four units of "ability" and he, only one unit—and they both are born with half of the wealth, $10,000 each. Andrea and Bob have the *same* von Neumann-Morgenstern utility function over ability and wealth, namely:

$$u(W, E) = E^{1/2}W^{1/2}$$

where E is units of ability and W is units of money wealth.

Because they have the same preferences, it turns out that the equilibrium in the market for contingent claims behind Dworkin's veil, where each soul assumes that it will receive an allotment of ability of one (four) with probability one-half (one-half), is characterized by the solution to the following optimization problem:

$$\operatorname*{Max}_{x} \frac{1}{2}u(10 + x, 1) + \frac{1}{2}u(10 - x, 4)$$

Let x^* be the solution to this program. Then the Dworkin tax scheme is that in which Andrea will transfer x^* thousand dollars to Bob.

The solution to this program is $x^* = -6$; in other words, *disabled Bob must transfer $6000 to able Andrea.*

Thus the Dworkin insurance market can produce the same pathology as afflicted the allocation chosen by the generalized Harsanyi IO in the last section. I contend that no resource-egalitarian would endorse this recommendation.

Unlike the situation in sections I and II, there is no interpersonal comparison of welfare required here. The utility function u represents a von Neumann-Morgenstern preference order over lotteries on prizes (W, E). Bob and Andrea have the same preference order. Nowhere need we posit that the welfare of Bob and Andrea is interpersonally comparable. Thus, to speak of "equalizing the welfares" of Bob and Andrea is meaningless, unless we postulate, in addition, some way of comparing their welfare levels. Thus, the example is perfectly consistent with Dworkin's view[10] that interpersonal comparisons of welfare are not available.

---

[9] "Equality of Talent"; *Theories of Distributive Justice*, chapter 7; and "Three Egalitarian Views and American Law."

One might protest that my example is cooked to produce the unpleasant[11] result. On the contrary, I contend that, were we to model preferences in our own world, complex as they are, we would surely encounter instances of this kind of transfer in the equilibrium of the market for contingent claims behind the Dworkinian veil of ignorance. The example is not bizarre; the "pathology" that it exhibits is, for all practical purposes, unavoidable.

Without level-comparable information on welfare across persons, we cannot speak of the "prioritarian" view, which says that we should give priority (in the distribution of wealth) to the worse-off.[12] We do not have such information in the Dworkin insurance scheme. But because the insurance mechanism ignores such level-comparable information, the allocation it recommends cannot generally favor the worse off. In particular, suppose we *do* have level-comparable information about welfare, and we believe that Bob with $10,000 is worse off than Andrea with $10,000 in the above example—not an outrageous view. Then clearly the insurance mechanism behaves in an antiprioritarian manner. Indeed, it is precisely a prioritarian intuition which lies behind my claim that distributive justice should render disabled persons more wealthy than able ones, other things equal.[13]

IV. CONCLUSION

We have extended Harsanyi's impartial observer argument by enriching the environment postulated by Harsanyi to include information on welfare levels of individuals. Thus, individuals are equipped both with preferences over wealth lotteries, and a conception of well-being which is interpersonally level comparable. We suggested a natural axiom, the principle of neutrality; together with Harsanyi's principle of acceptance, we are able to determine completely the preferences over lotteries over prospects of a representative impartial observer.

---

[10] Enunciated in Dworkin, "What Is Equality? Part 1: Equality of Welfare," *Philosophy and Public Affairs*, x (1981): 185-246.

[11] When I thus describe the result, I mean, from an egalitarian's viewpoint. Those who are persuaded that Dworkin's insurance mechanism is a compelling ethical procedure will not find the result unpleasant.

[12] On prioritarianism, see Derek Parfit, "Equality and Priority," *Ratio: New Series*, x (1997): 202-21.

[13] It is noteworthy that such an intuition need not be based on a welfarist view; it could be that we believe for reasons other than welfare, disabled people should have more material resources than able ones. This would follow, for example, from Sen's capability approach—*Commodities and Capabilities* (Amsterdam: North-Holland, 1985). See also the discussion of Tiny Tim, in G. A. Cohen, "On the Currency of Egalitarian Justice," *Ethics*, xcix (1989): 906-44

This enabled us to compute exactly what distribution of wealth would be recommended, from behind the veil of ignorance, by the impartial observer for the society that it represents.

We showed, in a series of examples with standard sorts of preferences, that the recommended distribution contravenes what most egalitarians would recommend, namely, that ceteris paribus, disabled individuals should receive more transferable resources (wealth) than able ones. Here, disability means, simply, the requirement of a greater wealth increment than others to achieve a given saltus in welfare.

Although the "egalitarian principle" here contravened may also be contravened by utilitarianism, it is important to note that the present argument is not utilitarian. As I noted earlier, utilitarianism is not even a well-defined concept, given the information postulated in our environment. Just because the veil of ignorance and utilitarianism may engender similar results does not mean they are the same thing.

Neither does the environment we have specified in section I permit us to study some of the more interesting questions that recent political philosophy has focused upon, such as how responsibility, circumstances, and effort impinge upon the requirements of justice. We are to think of the environment here postulated as one in which these issues do not arise. If we cannot solve the problem of justice in this simple environment, how are we ever to solve it in more complex environments? (But, in section III, where we study Dworkin's mechanism, responsibility does come into play, as we hold persons responsible for their preferences.)

Thus, in sections II and III, if a person is "disabled," it is here assumed to be through no fault of his own, or if a person has expensive tastes, it is, likewise, a characteristic for which we do not hold her responsible. Consequently, under these assumptions, most egalitarians would recommend that the "disabled," and persons with expensive tastes, receive more resources than the "able," and persons with cheap tastes, respectively. We have therefore illustrated what appears to be a fundamental inconsistency between egalitarianism and the recommendation following from a veil-of-ignorance construction.

Perhaps the conclusion should be that egalitarianism—or, more weakly, the egalitarian principle that I have invoked—is insupportable. To the contrary, I wish to propose that the veil-of-ignorance thought experiment is not a good one.

There is a cost and a benefit to using the veil of ignorance. We often mention the benefit, seldom the cost. Truth be told, it would be better to make decisions ex post, that is, after we know which preference orders and welfare functions are associated with which social positions (that is, after the birth lottery has transpired). The problem with making distributional decisions ex post is one of maintaining objectivity: How can we be sure that the decision makers, if they are drawn from the society in question, are not simply making recommendations from self-interest? The *benefit* of the veil-of-ignorance construct is that it forces objectivity, or impartiality. But the *cost* is that we must make decisions with a great handicap—we have discarded massively important information that is available to us in the real world, namely, what the actual joint distribution of resources (here, wealth and internal sources) and types *is*. The veil-of-ignorance approach asks us how we would allocate resources *if we did not know that actual distribution*. But would it not be better to think about the problem of distribution (now, *re*distribution) knowing what the actual distribution is, *if* we could otherwise maintain impartiality? The answer is surely yes, because we, or the decision maker in question, would have much more information available. Maximizing expected utility, ex ante, is a *pis-aller*, a technique a decision maker must use, should she be required to take an action before the state of the world has been revealed. But if that decision maker can afford to postpone her decision until the world's state has been revealed, *so much the better*. How, then, can we maintain the impartiality achieved, coercively, by the veil-of-ignorance thought experiment, but without incurring the substantial cost that accompanies it?

To be more specific: by forcing the decision maker to decide in the ex ante posture, we admit an element into the inquiry that has no obvious relevance to the question of distributive justice, namely, the element of a decision maker's preferences under uncertainty. I say such preferences have no obvious relevance to the question at hand, because they emerge as central *only* because of our method, that is, of constructing a veil of ignorance. It is certainly not a priori obvious why any individual's preferences over lotteries should influence what the ethically correct distribution of resources is. Allowing risk preferences to influence our decision about what distributive justice requires is a *cost* of the veil-of-ignorance method of inquiry—a cost, I say, we should seek to avoid, and perhaps can avoid, to bear. (In contrast, it is surely salient to allow a decision maker's preferences over lotteries to influence our evaluation of whether insurance markets are efficient. For uncertainty is of the essence when insurance is at issue.)

What I have argued is that that cost is manifest in precluding us from supporting a principle—that disabled individuals should, ceteris paribus, receive more resources than able ones. Faced with this preclusion, we should at least question the method that led to it.

We may state an analogy with the treatment of a disease. Perhaps there is a powerful medicine that will cure the disease, but at the cost of inducing a debilitating side effect. We may well decide not to use the medicine, and to seek another solution. I am saying that the veil of ignorance is a powerful medicine, but it requires us to sustain the side effect of employing preferences under uncertainty.

Although Dworkin's veil of ignorance is not as thick as Harsanyi's or Rawls's, his insurance mechanism is victim to the same problem that we have identified with the thick veil, namely, the possible (and likely) transfer of resources from disabled to able persons. Indeed, exactly the same critique applies to Dworkin's insurance scheme: it achieves impartiality at too high a cost.

I do not have an algorithmic answer to the question posed five paragraphs above, but the vague answer is that we must exercise our independent capacity for impartiality. As Serge-Christophe Kolm[14] has written, in a critique of original-position arguments: "justice is not blind-folded egoism, but open-eyed and informed objectivity" (*ibid.*, p. 20). Brian Barry[15] makes the distinction between contractarian arguments and arguments from impartiality, and I think he has the right idea. Thomas M. Scanlon[16] has the right idea when he tries to construct criteria for the admissibility of arguments that citizens can make to each other (we might call these *criteria of impartiality*). These authors are attempting to argue for varieties of egalitarianism without the tool of the veil of ignorance, by appealing directly to impartiality. What I hope to have shown is that this is the only way to provide foundations for egalitarian ethics.

JOHN E. ROEMER

Yale University

---

[14] *Modern Theories of Justice* (Cambridge: MIT, 1996).

[15] *Justice as Impartiality* (New York: Oxford, 1995).

[16] *What We Owe to Each Other* (Cambridge: Harvard, 1998).

INDUCTION, HUME, AND PROBABILITY*

David Hume argued, famously, that our reasonings about matters of contingent fact beyond the present evidence of our senses or the records of our memory (inductive reasonings in short) are not founded on any process of the understanding. Of the various sorts of attempt to defuse Hume's argument which have been made, one which many have found attractive is to appeal to some conception of objective probability.[1] But for each such conception that is appealed to, others have argued that Hume's argument cannot be defused by appeal to that conception.[2] Yet what has not been recognized, so far as I am aware, is that Hume himself makes a point about objective probability which entails that his argument cannot be defused by appeal to such a conception. His conception of objective probability was the classical one, but the point he makes, I aim to show, applies also to later conceptions, and even to a conception of subjective probability. In light of this point, Hume would have had no difficulty in combating attempts to defuse his argument by appeal to these conceptions.

Each of the conceptions of probability at issue faces well-known difficulties, but for the purpose in hand I shall set these aside, and focus only on that feature of each conception which enables Hume's point to get a grip.

According to one advocate of probability as a means of defusing Hume's argument, commentators on Hume have without exception failed to extract anything of philosophical interest from his discussion of probability, and with reason since this discussion is not only intrinsically unrewarding but also an inessential part of Hume's

---

*For valuable comments on an earlier draft of this essay, I am indebted to Lars Bergström, Gunnar Björnsson, Robert Callergård, and Paul Needham.

[1] Use of objective probability to provide an answer to Hume begins with Richard Price's application of the work of Thomas Bayes (though Price does not mention Hume by name) and continues down to the work of David C. Stove. See the Appendix, written by Price, to Bayes, "An Essay towards Solving a Problem in the Doctrine of Chances," *Philosophical Transactions of the Royal Society of London*, LIII (1763): 370-418, and Stove, *The Rationality of Induction* (New York: Oxford, 1986).

[2] See, for example, John Maynard Keynes, *A Treatise on Probability* (London: MacMillan, 1921), pp. 377-78; C. D. Broad, "The Relation between Induction and Probability, I-II," in *Induction, Probability, and Causation Selected Papers of C. D Broad* (Boston: Reidel, 1968), pp. 1-52; William Kneale, *Probability and Induction* (New York: Oxford, 1949), pp. 201-14.

0022-362X/02/9904/185–99

philosophy of induction.[3] Whether or not this judgment is correct where commentators on Hume are concerned, it could not be further from the truth where Hume's discussion of probability is concerned; for, as I aim to show, not only is this discussion essential for a proper understanding of Hume's philosophy of induction, but it is also of great interest in its own right.

I. HUME'S ARGUMENT AND THE CLASSICAL CONCEPTION OF PROBABILITY

Hume presents his argument both in the *Treatise* (Book I, part III, section VI) and the first *Enquiry* (section IV). The presentation in the first *Enquiry* represents Hume's last published thoughts on the matter, and differs in some important respects from the presentation in the *Treatise* (it does not, for example, presuppose Hume's analysis of causation). For this reason, I shall base my account of the argument on the presentation in the first *Enquiry*.

Hume is concerned with the nature of the evidence that assures us of a matter of fact beyond the present testimony of our senses or memory, and he considers whether it is through reasoning that we acquire this assurance. He[4] argues that all reasonings about matters of fact beyond the present evidence of our senses and memory are founded on the relation of cause and effect: it is by means of this relation alone that "we can go beyond the evidence of our memory and senses" (EN, p. 26 [4, 4]). When we are presented with an object (a loaf of bread, say) with certain sensible qualities (such as color, shape, and consistency), we can infer that the object will be *F* (nourishing, say) only given the premise that *F* is an effect of an object with these sensible qualities. But we can know that *F* is an effect of an object with these sensible qualities only through experience, and not through reason alone: presented with some object whose sensible qualities we have never before experienced in one object, we cannot infer solely from the sensible qualities of this object what effect it will have (or what its cause was); we can make this inference only if we have experienced the effect of objects with similar sensible qualities (or possess information on the matter that derives from someone who has had this experience) (EN, pp. 27-31 [4, 4-12]). Once we have acquired this experience, if we are presented with an object with similar sensible qualities, we infer that it will have a similar effect; from the judgment 'An object with such sensible qualities has always been attended with such an effect' we infer 'Other objects with similar sensible qualities will be attended with a similar effect' (EN, p. 34 [4, 16]). The connection between these two judgments, Hume continues, is not self-evident, and does not depend on any argument. All arguments are either demonstrative, having to do with relations of ideas, or moral (that is, probable), having to do with matters of fact. The two judgments cannot be connected by a demonstrative argument, because it implies no contradiction that an object with similar sensible qualities to those of objects we have experienced will not be attended with a similar effect; nor can the judgments be connected by a probable argument, because all probable arguments are premised on the assumption that objects with similar sensible qualities to those of objects we have experienced will have similar effects (the uniformity assumption), and hence a probable argument would be viciously circular (EN, pp. 34-37 [4, 17-22]). The two judgments, therefore, cannot be connected by any argument at all. Thus, Hume concludes, "in all reasonings from experience, there is a step taken by the mind which is not supported by any argument or process of the understanding" (EN, p. 41[5, 2]).[5]

---

[3] See Stove, *Probability and Hume's Inductive Scepticism* (New York: Oxford, 1973), Appendix, especially p. 120.

[4] *An Enquiry concerning Human Understanding*, L. A. Selby-Bigge, ed., third edition revised by P. H. Nidditch (New York: Oxford, 1975); *An Enquiry concerning Human Understanding*, Tom L. Beaucham, ed. (New York: Oxford, 2000). Hereafter EN, references in the text are to the pages of the former edition, with the section and paragraph numbers of the latter edition given in square brackets.

Some philosophers try to get round the need to invoke a causal premise in inductive reasoning by appealing to an objective relation of probability. On this line, inductive reasoning depends, not on the relation of cause and effect, but on an objective relation of probability $p(h|e) = r$ that obtains between the judgment e that reports the evidence of our senses and a judgment h that goes beyond that evidence, and since this relation does not depend upon the relation of cause and effect, Hume's argument does not get a grip.

The conception of probability with which Hume was familiar, the classical, has to do primarily with the outcomes of trials in games of chance, and hence is too narrow in scope to be used as a means of defusing Hume's argument. But it might be held that it can be used to defuse the argument at least where inductive reasoning about a matter of chance is concerned, for in this case there appears to be no need to assume that the matter of fact in question is the effect of any cause.

---

[5] My understanding of Hume's argument owes much to Lars Bergström's philosophical objections to the argument as I presented it in earlier drafts.

Hume himself discusses probability at length in the *Treatise*[6] (Book I, part III, section XI [Book I, part 3, section 11]) and more briefly in the first *Enquiry* (section VI). Since his treatment of the topic is substantially the same in both books, I shall take the fuller treatment in the *Treatise* as the basis of my discussion.

Chance, Hume writes, "is nothing real in itself, and, properly speaking, is merely the negation of a cause" (TR, p. 125 [11, 4]). By 'negation of a cause' I understand Hume to mean 'absence of any determining cause' rather than 'presence of some unknown determining cause'; but so far as the following discussion is concerned, nothing hinges on the question which of these two readings is correct.

Hume considers the case of a fair six-sided die, four sides of which have a certain mark (one spot, say), and two sides have a different mark (two spots). The classical probability (probability, for short) that the outcome of the next throw of this die will be a one-spot side is 2/3, and the probability that it will be a two-spot side is 1/3. In view of this, we expect a one-spot outcome on the next throw more than a two-spot outcome.

When we scrutinize what Hume says about probability, it becomes clear that Hume maintains, contrary to the view mentioned above, that the relation of probability does depend upon the relation of cause and effect. The following passage contains the essential clue to his thinking.

> And here 'tis remarkable, that tho' chance and causation be directly contrary, yet 'tis impossible for us to conceive this combination of chances, which is requisite to render one hazard superior to another, without supposing a mixture of causes among the chances, and a conjunction of necessity in some particulars, with a total indifference in others. Where nothing limits the chances, every notion, that the most extravagant fancy can form, is upon a footing of equality; nor can there be any circumstance to give one the advantage above another. Thus unless we allow, that there are some causes to make the dice fall, and preserve their form in their fall, and lie upon some one of their sides, we can form no calculation concerning the laws of hazard (TR, p. 125 [11, 6]).

The crucial point Hume makes here is that, in order to calculate the probability of a given outcome of a trial, we have to assume that there is a fixed finite number of possible outcomes; without this assumption, the calculation simply cannot be made, since the probability of the given outcome is the ratio between two fixed numbers, the number of possible outcomes of the given sort and the total number of possible outcomes. But this assumption is a *causal* assumption, for the sort of possibility that is involved is causal, not *logical*. In the case of the die, the fixed number of possible outcomes is determined by the six sides of the die; but these are not the only logically possible outcomes, for it is logically possible that the die, once thrown, flies off into the air, or explodes, or turns into a sphere, or lands on one of its twelve straight edges, or what you will. We assume that none of these things will happen, and that the die will land on one of its sides; we assume, in other words, that the *effect* of a throw is that the die lands on one of its six sides, though which side in particular may not be determined. This is what Hume means in the above passage by the phrases 'a mixture of causes among the chances' and 'a conjunction of necessity in some particulars, with a total indifference in others'. The causes in question are the force of gravity, the solidity of the die, its cubical form, and so on; these causes determine the die "to fall, to preserve its form in its fall, and to turn up one of its sides" (TR, p. 128 [11, 10]). Thus, contrary to the view in question, the need for a causal premise in inductive reasoning cannot be got round by invoking the relation of probability.

---

[6] *A Treatise of Human Nature*, 1739, L. A. Selby-Bigge, ed , second edition revised by P. H. Nidditch (New York: Oxford, 1978); *A Treatise of Human Nature*, David Fate Norton and Mary J. Norton, eds. (New York: Oxford, 2000). Hereafter TR; references in the text are to the pages of the former edition, with the section and paragraph numbers of the latter edition given in square brackets

As is the case with knowledge of any effect, we can know what the effect of a throw of a die will be only through experience and not through reason; if the effect is that the die lands on one of its sides, then we can know this only by experiencing it. Given this experience, we expect that an object that has similar sensible qualities to those of objects we have experienced will, if thrown, land on one of its sides. If this expectation could be supported by some argument, then the argument would rest on the uniformity assumption. Thus, the need for the uniformity assumption cannot be got round by appeal to probability.

As for the influence that a judgment of probability has on the mind, Hume acknowledges that the judgment 'A one-spot outcome has twice as many equal chances as a two-spot outcome' moves the mind to expect a one-spot outcome on the next throw more than a two-spot outcome, but he denies that it does so through grasp of any argument. He writes:

> Here we may repeat all the same arguments we employ'd in examining that belief, which arises from causes; and may prove after the same manner, that a superior number of chances produces our assent neither by *demonstration* nor *probability* (TR, p. 126 [11, 7]).

It cannot be demonstrated that the outcome will be the one that has the greater number of equal chances, for it is not necessary that this will be the outcome, and it is of no avail to argue that it is probable that this will be the outcome, for to say that it is probable that the outcome will be the one that has the greater number of equal chances is simply to say that the outcome with the greater number of equal chances has the greater number of equal chances (TR, p. 127 [11, 8]).

Hume appears to be treating probability as a purely descriptive relation, and to be conceiving of the probability of a given outcome as the ratio between the number of equal chances of the given outcome and the total number of equal chances or, to put this in more modern terms, the ratio between the number of possible outcomes of the given sort to the total number of equally possible outcomes. Given this purely descriptive conception of probability, a judgment of probability does not in itself imply anything about what we ought to believe or expect (this point calls to mind what Hume says about the relation between an *is* judgment and an *ought* judgment in the context of morality) (TR, p. 469 [1, 27]).

It might be objected that part of the content of a judgment of probability is normative: properly construed, the probability of an outcome of a given sort is the ratio between the number of possible outcomes of the given sort to the total number of equally credible possible outcomes, where by 'equally credible' is meant 'equally deserving of expectation'. Given this construal, a judgment of probability implies that we ought to expect each possible outcome to the same degree, and it is in virtue of this implication that a judgment of probability has an influence on the mind.

If the present objection had been put to Hume, he would surely have asked that it be explained under what conditions possible outcomes are equally deserving of expectation. These conditions, I believe, are purely causal.

Possible outcomes are equally deserving of expectation only if the standing factors are not such as to favor one possible outcome more than another, that is, are not such as to cause one possible outcome to turn up more frequently than another. In the case of the die, for example, the standing factors, such as gravity, the cubical shape of the die, the homogeneous composition of the die, and so on, do not favor one side more than another. But the condition that the standing factors are not such as to favor one possible outcome more than another is a causal condition, a negative one to be sure, but no less causal for that.

While the said condition is necessary if possible outcomes are to be equally deserving of expectation, it is not sufficient. Given the said condition, it is not the case that we ought to expect one possible outcome more than another, but it does not follow from this that we ought not to expect one possible outcome more than another, let alone that we ought to expect each possible outcome equally, for the judgment 'It is not the case that we ought to ϕ (to smoke, say)' does not entail 'We ought not to ϕ'. On the said condition, I cannot see that there is anything that we ought to expect or ought not to expect. It will not do to say that, on the said condition, each possible outcome will occur as frequently as any other and this fact gives us reason to expect each possible outcome equally, for the proposition that the standing factors are not such as to cause one possible outcome to occur more frequently than another does not entail that each possible outcome will occur as frequently as any other. Nor will it do to say that as the number of trials goes to infinity the probability goes to 1 that each possible outcome will occur as frequently as any other, for to appeal to a judgment of probability as grounds for taking the said condition to be sufficient would be to beg the question.

A further condition is needed, therefore, if possible outcomes are to be equally deserving of expectation. The requisite condition is that the standing factors are such as to favor each possible outcome equally, that is, are such as to cause each possible outcome to occur as frequently as any other. But, once again, this condition is causal.

Any argument, therefore, about what we ought to expect that is premised on a judgment of probability rests upon a causal assumption, and hence on the uniformity assumption.

In sum, not only is a causal assumption required if a judgment of probability is to be made in the first place, but a further causal assumption is required if the judgment is to have normative force. It is not on account of any argument, therefore, that a judgment of probability has normative force, and thereby has an influence on the mind, for any such argument must be premised on the uniformity assumption.

II. HUME'S ARGUMENT AND THE LOGICAL CONCEPTION OF PROBABILITY

Those philosophers who look to probability as a means of disarming Hume's argument need not initially be put out by the considerations of the previous section, for they can appeal to conceptions of probability, such as the logical conception, which were developed after Hume's time, and which appear to be more suited to the task. At first glance, the logical conception appears to leave no foothold for Hume's argument; for it represents the relation $p(h|e)$ as a purely logical relation between the propositions h and e which depends only on the meaning of the sentences 'h' and 'e', and not on any causal or uniformity assumption.[7]

On the logical conception, the relation of probability purports to be a generalization of the relation of entailment. Given the semantic conception of entailment, what it is for a proposition e to entail a proposition h is for h to be true in all the possible worlds in which e is true, the possibility in question being logical. By analogy, what it is for the probability of h given e to be r is for h to be true in the proportion r of possible worlds in which e is true.

The most thoroughgoing attempt to develop the logical conception of probability is that of Rudolf Carnap in *Logical Foundations of Probability* (*ibid.*). Although Carnap later abandoned the theory he presented in that book, it is of interest to consider the theory from Hume's angle.

In Carnap's theory, the key idea is that of a state-description in a formal language. A state-description is a sentence in a formal language L that gives an exhaustive description of one possible world describable in L by stating, for every individual constant a and every predicate F in L, whether or not a belongs to F (*ibid.*, p. 72). For example, in the language L_3 which contains three individual constants a, b, c and one predicate F, there are eight state-descriptions: $(Fa \mathbin{\&} Fb \mathbin{\&} Fc)$, $(Fa \mathbin{\&} Fb \mathbin{\&} \neg Fc)$, $(Fa \mathbin{\&} \neg Fb \mathbin{\&} Fc)$, $(Fa \mathbin{\&} \neg Fb \mathbin{\&} \neg Fc)$, $(\neg Fa \mathbin{\&} Fb \mathbin{\&} Fc)$, $(\neg Fa \mathbin{\&} Fb \mathbin{\&} \neg Fc)$, $(\neg Fa \mathbin{\&} \neg Fb \mathbin{\&} Fc)$, $(\neg Fa \mathbin{\&} \neg Fb \mathbin{\&} \neg Fc)$. There is a measure function m for the state-descriptions which assigns a positive real number to each state-description, the values of m for all state-descriptions summing to 1 (the values of m are commonly referred to as "weights," though not by Carnap, who speaks only of "values"). The function m is extended to every sentence s of L, so that $m(s)$ is the sum of the weights of the state-descriptions containing s. On the basis of this apparatus, the probability c of h given e is defined as the ratio of the sum m of the weights of the state-descriptions in which h and e are both true to the sum m of the weights of the state-descriptions in which e is true, that is, for any pair of sentences h, e in L where $m(e) > 0$, the following equation holds:

$$(1) \quad c(h|e) = \frac{m(h \& e)}{m(e)}$$

What remains is the choice of function m for the state-descriptions. The choice that springs readily to mind is the function $m\dagger$ (Carnap's symbol) which assigns the same weight $1/\nu$ to each state-description in L, where ν is the number of state-descriptions in L. But as Carnap observed, on this choice, if h is the sentence 'Fa_{k+1}' ('The individual experienced after the k^{th} will be F'), and e is the sentence 'Fa_1 & Fa_2 &...& Fa_k' ('All the k individuals experienced were F'), then the probability of h given e is $1/2$ for all values of k less than n. Thus, given that all the k loaves of bread we have experienced were nourishing, the probability that the $(k + 1)^{\text{th}}$ loaf we shall experience will be nourishing is $1/2$ for any value of k less than n. Moreover, if e' is formed from e by replacing some or even all the atomic sentences in e with their negations, the probability of h given e' remains $1/2$ (*ibid.*, pp. 564-65).

---

[7] See Rudolf Carnap, *Logical Foundations of Probability* (Chicago: University Press, 1950), p. 181.

Carnap's solution to the above problem was to stipulate that what are to be assigned equal weight are not the state-descriptions but the structure-descriptions, a structure-description being a sentence in L that specifies the way in which the predicates in L are instantiated. In L_3, for example, there are four structure-descriptions: $[(Fa \,\&\, Fb \,\&\, Fc)]$, $[(Fa \,\&\, Fb \,\&\, \neg Fc) \vee (Fa \,\&\, \neg Fb \,\&\, Fc) \vee (\neg Fa \,\&\, Fb \,\&\, Fc)]$, $[(Fa \,\&\, \neg Fb \,\&\, \neg Fc) \vee (\neg Fa \,\&\, Fb \,\&\, \neg Fc) \vee (\neg Fa \,\&\, \neg Fb \,\&\, Fc)]$, and $[(\neg Fa \,\&\, \neg Fb \,\&\, \neg Fc)]$. Each structure-description is assigned the same weight $1/\tau$ by the measure function m^*, where τ is the number of structure-descriptions in L. The weight $1/\tau$ is then divided equally among the state-descriptions contained in the structure-description, so that each state-description has the weight $1/\tau\zeta$, where ζ is the number of state-descriptions in the structure-description in question. Thus the state-descriptions no longer have equal weight, since the value of $1/\tau\zeta$ varies with the value of ζ (*ibid.*, pp. 563-64). In L_3, for example, the weight of each of the two state-descriptions $(Fa \,\&\, Fb \,\&\, Fc)$ and $(\neg Fa \,\&\, \neg Fb \,\&\, \neg Fc)$ is $1/4$, whereas the weight of each of the remaining six state-descriptions is $1/4 \div 3 = 1/12$.

Given this new basis, Carnap defined the probability c^* of h given e as the ratio of the sum m^* of the weights of the state-descriptions containing h and e to the sum m^* of the weights of the state-descriptions containing e; in other words, for any pair of sentences h, e in L where $m^*(e) > 0$

$$(2) \quad c^*(h|e) = \frac{m^*(h \,\&\, e)}{m^*(e)}$$

On this definition, the probability of h given e rises above $1/2$ as the number of individual constants in e increases, provided h is not a universal generalization. (If h is a universal generalization, then, as the number n of individual constants in L_n increases, the probability of h given e decreases, tending toward 0 as n goes to infinity. Jaakko

Hintikka[8] proposed a solution to this difficulty, but I shall not discuss it, since the difficulty is not a consequence of any point of Hume's, and in any case Hume was concerned with singular matters of fact, rather than universal matters of fact).

What, then, would Hume have had to say about Carnap's construction of c^*? When we scrutinize the construction through Hume's eyes, as it were, what strikes us is that the state-descriptions that have more weight are exactly those which describe *uniform* possible worlds. In L_3, for example, the state-descriptions that have the most weight are $(Fa \,\&\, Fb \,\&\, Fc)$ and $(\neg Fa \,\&\, \neg Fb \,\&\, \neg Fc)$. Each of these uniform state-descriptions has three times the weight of any of the other nonuniform ones. A bias toward uniformity is thus built into Carnap's theory.

There would be no cause for complaint if the bias toward uniformity in Carnap's theory were a consequence of some compelling, purely logical consideration; indeed, there would be cause for congratulation, for the uniformity assumption would have been given a basis in logic. But this is not the case, for the bias is a consequence of Carnap's stipulation that the structure-descriptions are to have equal weight, and this stipulation is not based on any compelling logical consideration. On the contrary, the stipulation was made solely for the purpose of constructing a numerical function for probability which was not at odds with our inductive intuitions. Carnap admitted this himself, and acknowledged that this is not a strong reason for the stipulation (*op. cit.*, pp. 564-65). This is an understatement, for structure-descriptions vary considerably in their structural features (as Carnap also recognized): some structure-descriptions contain only one state-description, while others contain several. If the weight of a structure-description is a measure of its chances of being true of the actual world, and its chances of being true of the actual world are the number of distinct ways in which it could be true of the actual world, then some structure-descriptions have more chances than others, and hence should have more weight. State-descriptions, by contrast, do not vary in any way that affects their chances of being true of the actual world, and hence ought to be assigned equal weight. But to assign equal weight to the state-descriptions would be to choose $m\dagger$, and thereby also $c\dagger$, and this choice, Carnap states, "would be tantamount to the principle never to let our past experiences influence our expectations for the future" (*op. cit.*, p. 565). Indeed; but the lesson to be learned from this, Hume would say, is not that the state-descriptions should be assigned unequal weight, but that it is not through grasp of any logical relation that our past experiences influence our expectations for the future. The bias toward uniformity that is built into Carnap's logic, therefore, is ad hoc.

---

[8] "A Two-dimensional Continuum of Inductive Methods," in Hintikka and Patrick Suppes, eds., *Aspects of Inductive Logic* (Amsterdam: North-Holland, 1966), pp. 113-32.

III HUME'S ARGUMENT AND THE FREQUENCY AND PROPENSITY CONCEPTIONS OF PROBABILITY

As for the frequency and propensity conceptions of probability, little effort is required to recognize that these offer no prospects of defusing Hume's argument.

On the frequency conception, the probability of an individual *F* being *G* is the limit *r* of the relative frequency of *G*s in an indefinitely long random sequence of *F*s (von Mises's definition). That there is a limit is something we can know only through experience, and not through reason, since the limit is not determined by any rule; if there is a limit, we can know what it is only by inductive inference from the fact that, in a long sequence of *F*s of which we have experience, the relative frequency of *G*s, after a certain position in the sequence, does not vary from the value *r*. Thus, any attempt to defuse Hume's argument by means of the frequency conception will be circular. Richard Von Mises[9] was well aware that knowledge of the limit is by inductive inference, though he did not employ the expression 'inductive inference', and appeared not to be concerned with the nature of this inference. By appealing to certain pragmatic considerations, Hans Reichenbach[10] argued that the obstacle of circularity could be overcome, but few have found his argument persuasive.

As for the propensity conception, this represents the probability of an event of a given kind as a measure of the strength of the propensity of an object, or object plus setup, of a certain kind to give rise to an event of that kind.[11] Thus conceived, a propensity is a causal item, since the expression 'to give rise to' is tantamount to 'to cause'. A propensity is similar to what Hume calls a 'power', and his argument applies just as much to powers as it does to effects.[12] We can know that an object whose sensible qualities we have never before encountered in one object has a certain power only through experience, and not through reason: given this experience, when we encounter a new object that has similar sensible qualities, we expect that the new object will have a similar power. We can know, for example, that an object that has the sensible qualities of a die has a propensity to land with a given side uppermost only through experience of the relative frequency with which objects with these sensible qualities have landed with that side uppermost; given this experience, when we encounter a new object with the sensible qualities of a die, we expect it to have a similar propensity. If this expectation were based on reasoning, then the reasoning would rest on the premise that objects with similar sensible qualities to those we have experienced have similar propensities, which is a variant of the uniformity assumption.

---

[9] *Probability, Statistics, and Truth* (New York: Dover, 1981, second edition), p. 15.

[10] *Experience and Prediction: An Analysis of the Foundations and Structure of Knowledge* (Chicago: University Press, 1938), p. 313. Wesley Salmon adopted Reichenbach's approach in *The Foundations of Scientific Inference* (Pittsburgh. University Press, 1966).

[11] See Karl Popper, "The Propensity Interpretation of the Calculus of Probability, and the Quantum Theory," in Stephan Körner, ed., *Observation and Interpretation*, (London: Butterworth, 1957), pp. 65-70.

[12] See Hume, EN, p. 36 [4, 21].

There is, so far as I can judge, no prospect of defusing Hume's argument by appeal to objective probability.

IV HUME'S ARGUMENT AND THE BAYESIAN CONCEPTION OF PROBABILITY

The question remains whether Hume's argument can be defused by dispensing with objective probability altogether and adopting instead the strict version of the Bayesian conception, which admits of no appeal to objective probability.[13] On this version, probability is a measure p_A of the degree of belief of a given person A, and this degree of belief is rational if and only if p_A satisfies the axioms of the calculus of probability (for the sake of simplicity I write 'p' below rather than 'p_A').[14] On this conception the function $p(h|e)$ represents A's degree of belief in h on the assumption that e is true, and this degree is rational if and only if $p(h|e)$ satisfies Bayes's theorem given $p(e) > 0$:

$$(3) \quad p(h|e) = \frac{p(h) \cdot p(e|h)}{p(e)}$$

Various arguments for the said condition of rationality have been given, the most influential being the so-called Dutch book argument. According to one version of the argument, if A's degrees of belief do not satisfy the said condition, A can be presented with a set of bets which A judges to be fair even though, if A accepts these bets, she will make a net loss whether or not the propositions that are the subjects of the bets turn out to be true, which indicates that the degrees of belief on which this judgment is based are not rational; in order to be rational, it is necessary (and in fact sufficient) that A's degrees of belief satisfy the calculus of probability.

---

[13] For a recent attempt to answer Hume by appeal to a Bayesian conception, see Samir Okasha, "What Did Hume Really Show about Induction?" *The Philosophical Quarterly*, LI, 204 (2001): 307-27. In my mind, Okasha does not succeed, for reasons which I do not have the space to go into here.

[14] See F P. Ramsey, "Truth and Probability," in his *The Foundations of Mathematics and Other Logical Essays* (New York: Harcourt Brace, 1931), pp. 156-98; and Bruno de Finetti, "La prévision: ses lois logiques, ses sources subjectives," *Annales de l'Institut Henri Poincaré*, VII (1937): 1-68.

If the present conception is to offer a way of defusing Hume's argument, then it has to be shown, on the basis of the conception, that A's degree of belief in h rationally increases as A's evidence increases provided A encounters no evidence to the contrary. The only way in which this can be done, so far as I am aware, is to adopt the so-called *conditionalization condition*, which is to the effect that, if A comes to learn that e is the case, and A revises her degree of belief in h in light of this knowledge alone, then the revision is rational if and only if A's revised degree of belief in h is equal to $p(h|e)$. The standard argument for this condition is a Dutch book argument.[15]

Provided the conditionalization condition is satisfied, it can be shown that, with a probability of 1, A's degree of belief in h increases as A's evidence increases, and goes to 1 as the quantity of evidence goes to ∞. Haim Gaifman and Marc Snir[16] have proved this for a certain formal language. To paraphrase Gaifman and Snir's theorem: given certain conditions, the probability is 1 that the value of $p(h|e_0 \,\&\, e_1 \,\&\ldots\&\, e_{r<n})$ goes to 1 as n goes to ∞ if h is true and to 0 if h is false (*ibid.*, p. 507). Thus, if A revises her degree of belief in the hypothesis Fa_{r+1} in the light of each piece of new evidence Fa_0, $Fa_1,\ldots,Fa_{r<n}$, and the conditionalization condition is satisfied, then the probability is 1 that A's degree of belief in Fa_{r+1} goes to 1 if Fa_{r+1} is true, and to 0 if Fa_{r+1} is false as n goes to ∞.

We have here a model of degree of belief revision that might be thought to provide a way of defusing Hume's argument. For, if the process through which we are assured of a matter of fact beyond the present evidence of our senses fits the model, then the process is shown to be, though not a process of the understanding, at any rate a process that satisfies a condition (conditionalization) laid down by the understanding. But Hume's argument can be defused in this way only if the argument for the said condition does not call for the uniformity assumption.

---

[15] See Paul Teller, "Conditionalization and Observation," *Synthese*, XXVI (1973): 218-58.

[16] "Probabilities over Rich Languages, Testing, and Randomness," *Journal of Symbolic Logic*, XLVII, 3 (1982): 495-548. Other aspects of this theorem are discussed by John Earman in *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory* (Cambridge: MIT, 1992), pp. 147-49.

So far as the Dutch book argument for the said condition is concerned, this does not appear at first glance to involve any appeal to the uniformity assumption, but on closer inspection it becomes clear that it does. For the proposition that *A* will make a net loss is a proposition about the *effect* of such a bet on *A* and, as Hume maintains, this proposition, like all propositions about effects, can be known only through experience, and not through reason alone. Other effects of such a bet are conceivable. *A* may not hand over her stake to *B* (*A*'s betting partner), or *B*, having already received *A*'s stake, may give it back to *A*. In either case, *A* will not make a loss. We assume that *A* will hand over the stake to *B* and that *B* will retain it, but this assumption is based on our past experience that persons who made bets with each other fulfilled them, and in order to go beyond this experience, as Hume maintains, we need the uniformity assumption. The proposition '*A* will make a net loss' is thus in the same boat as the proposition 'The die will land on one of its sides'.

Whether there is another argument for the conditionalization condition which avoids appeal to the uniformity assumption I shall not consider; for there is a problem with the conditionalization condition itself, irrespective of the argument for it, that is of interest in the present connection. The problem is that, if the value of $p(h)$ on the right-hand side of equation (3) above is 0, then the value of $p(h|e)$ will be 0, and consequently the value of $p(h)$ will not rise as the quantity of evidence e increases. In other words, if *A*'s degree of belief in h is zero to begin with, that is, before *A* acquires any evidence relevant to the truth of h, then it will remain zero, no matter how great the quantity of evidence *A* comes to acquire thereafter. If h is a contingent proposition, then there is nothing in the model to say that it is not rational for *A* to have zero degree of belief in h to begin with. This problem is well known, but it is highly relevant in the present context, for it would be of no surprise to Hume. Hume's response would surely be that, if h is a proposition about a matter of fact beyond the present evidence of *A*'s senses and memory, then there can be no argument to convince *A* that her degree of belief should not be zero to begin with. So far as the understanding is concerned, one degree of belief in h to begin with is as good or as bad as any other. Hence, if *A*'s degree of belief in h does increase when she receives the evidence e, then it is not at the behest of *A*'s understanding that it does so.

It may be objected that *A* has more to go on to begin with than the understanding alone. Before *A* acquires evidence relevant to the truth or falsity of h, *A*'s understanding does not operate in a vacuum but within the set of *A*'s other beliefs (call this set 'b'), and in light of

b, A's degree of belief in h may well be greater than zero, that is, the value of $p(h|b)$ may be greater than 0. This is true, but of no avail, for on the present model the value of $p(h|b)$ can *rationally* be greater than 0 only if it satisfies Bayes's theorem, as follows given $p(b) > 0$:

$$(4) \quad p(h|b) = \frac{p(h) \cdot p(b|h)}{p(b)}$$

Now, given (4), the value of $p(h|b)$ can be greater than 0 only if the value of $p(h)$ on the right-hand side of (4) is greater than 0. So we are no further ahead.

If it is not rational for A's degree of belief in h to be zero to begin with, then I do not know of any argument to show this that is not ad hoc or does not beg the question against Hume.

To conclude: by invoking only points that he made in his discussion of probability or in his own original argument, Hume would have been able to defend his argument against attempts to defuse it by appeal to those conceptions of probability which have been developed since his time.

DUGALD MURDOCH

University of Stockholm

COMMENTS AND CRITICISM

AMAZING KNOWLEDGE

One of the goals of stem-cell research is supposed to be to enable us to learn how stem cells divide to become specialized cells. But we already have this knowledge. Let me explain.

Assume that knowing is a relation to Russellian propositions, structured entities whose basic components are the objects and properties our knowledge is about. Knowing is either a *triadic* relation that holds among a knower, a Russellian proposition the knower knows, and a mode of presentation under which the knower knows the proposition, or else it is a *dyadic* relation that holds between a knower and the Russellian proposition she knows, where this dyadic relation is explicable in terms of a triadic relation, KNO(x, p, m), which behaves just as knowing would behave if it were a triadic relation. In either case, we may speak of someone's knowing a proposition under one mode of presentation but not under another. Lois Lane, for example, knows the proposition that Superman/Clark Kent flies under a mode of presentation that requires thinking of Superman/Clark Kent as a guy who goes around in a caped spandex outfit but not under a mode of presentation that requires thinking of him as a bespectacled nerd.

Some modes of presentation are not at all like the one just invoked in the Lois Lane example; they are *practical* modes of presentation not easily put into words but ones with which a person may be *en rapport* in certain practical ways. For example, if w is a way in which a mechanism works, and that mechanism is at work in you, then there is ipso facto a practical mode of presentation m of the proposition that w is a way in which the mechanism works such that you know, under m, that w is a way in which the mechanism works.

Now, let w_s be the way in which stem cells divide to become specialized cells. It follows, then, that for each and every one of us there is some practical mode of presentation m such that one knows, under m, that w_s is a way stem cells divide to become specialized cells.

You are right to doubt that we know any such thing. If asked to justify your skepticism, you would no doubt appeal to the following facts:

(a) Typically, if a person knows p—and thus knows it under some mode of presentation or other—she is also able to assert p, where this entails the ability assertively to utter a sentence that, in the context

0022-362X/02/9904/200-02

of utterance, expresses *p*. Thus, although Lois Lane is not in a position to say 'Clark Kent flies', she is in a position to say 'Superman flies'. But none of us can produce a sentence that expresses the proposition that w_s is a way stem cells divide to become specialized cells.

(b) We are not given the slightest clue as to what "practical modes of presentation" are like.

(c) If you know that such-and-such, then someone who has both knowledge of all the facts relevant to your knowing that such-and-such and mastery of our concept of knowledge-that should recognize that you know that such-and-such. To imagine otherwise is sort of like imagining that there might be horses such that no one could know them to be horses however much one knew about those creatures. Yet it is clear that however much you learn about the situation of each and every person with respect to stem-cell specialization, no one with mastery of our concept of knowledge-that would credit any of us with knowledge that w_s is a way in which stem cells divide to become specialized cells. That is simply not how we deploy 'knows that'. Indeed, one who uses 'knows that' with its ordinary meaning would, given the stipulated facts, *deny* that anyone knows how stem cells divide to become specialized cells; nothing in the situation resembles anything that would ordinarily be taken as evidence for the amazing knowledge claim.

Mozart wrote the symphony K. 16 when he was eight years old. Clyde, a chicken-sexer, is able to look at a one-day old chick and straightway know its sex, although he says he has no idea how he is able to do that. Let w_m be the way in which eight-year-old Mozart wrote K. 16, and let w_c be the way Clyde is able to tell the sex of a chick. Now consider the claim that there are "practical modes of presentation" m and m' such that eight-year-old Mozart knew, under m, that w_m was a way to write K. 16 and Clyde knows, under m', that w_c is a way to discern the sex of a chick. The reasons just displayed for doubting we have the knowledge in question about stem cells also apply to the knowledge claims about Mozart and Clyde. I conclude that eight-year-old Mozart no more knew that w_m was a way to write K. 16 and Clyde no more knows that w_c is a way to discern the sex of a chick than we know that w_s is a way stem cells divide to become specialized cells.

If I am right about Mozart and Clyde, then Jason Stanley and Timothy Williamson are wrong when, in "Knowing How,"[1] they claim that knowledge-how is a species of knowledge-that. For they accept the account of knowledge sketched above and they claim that, if x knows how to Φ, then there is a practical mode of presentation m and a way w of Φing such that x knows, under m, that w is a way to Φ. And eight-year-old Mozart certainly knew how to write K. 16, and Clyde knows how to tell the sex of a one-day old chick.

---

[1] This JOURNAL, XCVIII, 8 (August 2001): 411-44.

Stanley and Williamson cannot escape attributing the propositional knowledge in question to Mozart and to Clyde; nor can they reasonably deny that the displayed reasons for doubting we have the mooted propositional knowledge about stem cells also apply to the knowledge claims about Mozart and Clyde. At the same time, the stem-cell case does not involve knowledge-how, whereas Mozart knew how to write K. 16 and Clyde knows how to tell the sex of a one-day old chick, and Stanley and Williamson have adduced some linguistic evidence in support of the claim that ascriptions of knowledge-how are ascriptions of propositional knowledge. Well, we have to weigh their linguistic evidence against the fact that the reasons for doubting that we know that w_s is a way in which stem cells divide to become specialized cells apply as well to the claim that eight-year-old Mozart knew that w_m was a way to write the symphony K. 16 and to the claim that Clyde knows that w_c is a way to discern the sex of a chick. I dare say, my own bookkeeping is unfavorable to Stanley and Williamson.

STEPHEN SCHIFFER

New York University

BOOK REVIEWS

Aristotle's Theory of Language and Meaning. DEBORAH K. W. MODRAK. New York: Cambridge University Press, 2001. ix + 302 p. Cloth $59.95.

In *De Interpretatione* 1, Aristotle presents what many have regarded as a general account of meaning, in the following terms:

> Spoken words then are symbols of affections in the soul and written words are symbols of spoken words. And just as written letters are not the same for all humans neither are spoken words. But what these primarily are signs of, the affections of the soul, are the same for all, as also are those things of which our affections are likenesses. This matter has, however, been discussed in my treatise on the soul...(16a3-9; translation follows Modrak).

When we turn to the *De Anima*, however, to which the above passage apparently directs us, we find very little that explicitly refers to language or meaning, as such, or even any indication that the content of the *De Interpretatione* passage is being filled in at all.

Despite this lack of explicit detail, many commentators have plausibly thought that the relevant discussion is that concerning imagination and thought to be found in book III, chapters 3-8 of the *De Anima*, and that the affections (*pathemata*) of the soul referred to in *De Interpretatione* 1 must be sensory images (*phantasmata*). As a result, they have tended to credit Aristotle with a somewhat primitive and naïve image theory of meaning—according to which the meanings of words are to be found in the mental representations of external objects, with which those words are associated—a view that is subject to numerous objections.

Deborah Modrak's new book, *Aristotle's Theory of Language and Meaning*, does much to dispel this rather disparaging assessment. She argues that through a consideration of the role that Aristotle assigns to the imagination in a number of cognitive processes, we can discern the psychological basis of a much more sophisticated image theory of meaning than that which is usually attributed to Aristotle. But in addition to this, she shows how Aristotle's general theory of meaning coheres with his philosophy of language, epistemology, and metaphysics. What results is an overall picture which is systematically cohesive and consistent, at least in its broad strokes, and which involves a theory of meaning that is more plausible than its less sophisticated alternative.

0022-362X/02/9904/203-09

The book is divided into three parts dealing, respectively, with the "epistemic, ontological, and psychological foundations of Aristotle's philosophy of language" (50). In part I (Language and Knowledge) Modrak spells out the general theory of meaning in the broader context of Aristotle's philosophy of language and epistemology, while in part II (Definition and Essence) she articulates the metaphysical presuppositions of his conception of language and epistemology. Both of these parts present fairly standard interpretations of important components of Aristotle's philosophy of language, epistemology, and metaphysics, or at least recognized alternative interpretations, and do so clearly and in accessible terminology. While Modrak contributes some useful insights for our understanding of these components, what is perhaps of greater interest is the way in which she ties these components together through the theme of Aristotle's theory of meaning. Part III (Cognition and Meaning) turns to consider the "psychological component of meaning," and it is here that Modrak offers her more original interpretive insights, explaining how Aristotle's account of imagination enables linguistic content that is general in its nature to be conveyed by images that are particular in their nature.

The overall idea of Aristotle's account, as Modrak presents it, is fairly straightforward—as long as we are willing to grant the presuppositions that are built into it: words have meaning through being associated with affections of the soul that are themselves likenesses, or representations of external things—both particular and universal. Those affections, Modrak plausibly argues, are intended by Aristotle to be individual *phantasmata* (images).[1]

This account presupposes a metaphysics and ontology according to which there are not only individual substances present in the world but also real universals—forms or essences realized in matter—and it presupposes an epistemology according to which those forms or essences are ultimately accessible to human minds. Such forms are the basic objects of scientific inquiry, and the corresponding universal concepts the "building blocks of language and knowledge."

There is, then, a parallel between language and thought, on the one hand, and the external world, on the other, whereby the structures manifest in our universal concepts represent and are isomorphic with the fundamental structures of reality. And the central problem for this overall conception, according to Modrak, is that of explaining how an individual *phantasma* can be both a sensory representation of a particular individual *and* a universal cognitive content:

---

[1] Of course, by 'images' Modrak does not mean just visual images, but complexes of visual, auditory, olfactory, gustatory, and tactile images—sensory images in general.

> Herein lies the crux of the problem for Aristotle's theory of meaning and the postulated relation between a *phantasma* and a meaning. *Phantasmata* seem consigned to the role of representing particulars, but meanings are universals (248).

The first step in resolving this problem involves the recognition of an important feature of Aristotle's epistemology. According to Aristotle, the acquisition of the universal concepts of physics and metaphysics is grounded in the process of induction (*epagoge*) described in *Posterior Analytics* II, 19 (Modrak argues that the acquisition of abstract concepts of mathematics and geometry are also empirically based, for Aristotle). Through the experience of various particulars of the same type, the universal form definitive of the type comes to rest in the soul. But this is a process that takes place to a greater or lesser degree, depending on the extent of one's experiences, and gives rise to concepts of greater or lesser generality, depending on the relative place of the type within an ontological hierarchy of types (human, animal, life-form, for example). Thus, according to Modrak, Aristotle:

> ...envisages a nesting of concepts where the experiences that determine the linguistic meaning also present the essence that is subsequently articulated in the scientific definition (277).

There are two related points here. The first, to which I shall return when considering the second stage of the solution, is that the process of induction gives rise to a nesting of concepts. The second is that the rudimentary universals of experience are the linguistic meanings of natural language, while scientific definitions of essence refine and correct these.

This latter observation helps explain how different users of the same language may associate nonequivalent concepts with the same words, and yet successfully communicate—for the more accurate and refined concepts of the scientific community are already "present in" the less refined concepts of the everyday person on the street. But the observation also assumes that the essences the scientist ultimately discovers are the same as those universals to which our everyday experiences converge. While this is an assumption with which Aristotle was apparently quite comfortable, it is not one that modern philosophers are likely to make without serious reservations. (Modrak is conscious of this concern (236), but does not give a satisfactory response to it.)

The central concern of the book, however, is with Aristotle's theory of meaning, and Modrak's account of the "psychological component of meaning" in part III completes and renders plausible the general account of meaning given by Aristotle in *De Interpretatione* 1. Her main insight is based on a careful reading and construal of the texts, and is supported by a consideration of Aristotle's account of a number of other cognitive functions—memory, thinking, the formation of abstract concepts, and the drawing of inferences—all of which involve imagination. Modrak's insight is that, according to Aristotle, the same sensory representation (a particular *phantasma*) can be used to "focus" on different levels of nested generality. The relevant *pathema*, then—the affection of the soul—is a particular *phantasma* representing a sensory individual, but employed by the rational faculty to think, or mean, a universal concept. This is possible because the content of the particular sensory representation (the *phantasma*) carries with it the content of the universal concept, and the noetic or linguistic act apprehends that universal concept, ignoring what is idiosyncratic to the individual itself.

Thus, for example, my image of Socrates may be used to think of, or mean, the individual Socrates, the species human, or the genus animal—and perhaps other more, or less general concepts—because the content of these progressively more general concepts (as Modrak says, "in the favored cases"—my experience might be very limited) is present in the sensory content presented in the image. This is a consequence of Aristotle's theory of knowledge noted above, according to which we acquire concepts more, or less general, as a result of our experience of objects falling within the same hierarchy of ontological types.

It is important to realize that what is being suggested is not that we have a (Lockean) abstract general *phantasma*, nor is it that we use a particular *phantasma* to stand proxy for all the individuals of the same type (as Berkeley might have suggested). Rather, it is that the universal is "present in" the individual *phantasma*, and that we use such a *phantasma* to grasp, comprehend, conceptualize, and *mean* the universal.

There are many more details to be filled in—and Modrak, of course, provides those details—but I see two related difficulties with this overall interpretation as it presently stands. The first is that it apparently has the noetic act, and the act of meaning or understanding something by a sign, as involving an *additional* act of cognitive discrimination. The mind, as it were, "sees" the universal in the *phantasma*, but unless we are very careful (and I am not sure that Modrak has quite taken enough care on this point), we are perilously close to the idea that the mind already knows what it is looking for.[2] The mind must think the universal to itself prior to (and independently of!) thinking the universal in the *phantasma*.

Only once does Modrak come close to disabusing us of this conception:

> The cognitive process involved in the acquisition of language and in the apprehension of essences is characteristic of human minds...the process is not driven by a conscious reinterpretation of the sensory representation; rather, it is a process that will take place given appropriate conditions of experience, training, intellectual activity, and historical situation, just as perception is a cognitive process that will take place under appropriate conditions without conscious effort on the part of the percipient. Humans are by nature language users (258).

But the conception needs to be resisted more explicitly than this, unless a wholly different level of cognitive activity is to be posited, and this does not seem to be the intention either of Aristotle, or of Modrak in her interpretation of his position.

My second concern is that the content of the *phantasma*, insofar as it is used as a representation of a universal, and the content of the (same) *phantasma*, insofar as it is a representation of an individual, are nonidentical contents (even though the one is nested within the other). Thus, Modrak states that:

> Under ideal circumstances, using the *phantasma* of a concrete particular as a representation of a token of a certain type enables the thinker to recognize the essence that is characteristic of the type. In such cases, the *pathema* qua *phantasma* is like (resembles$_1$) a token of the type, and the *pathema* qua meaning is also properly described as a likeness (a resemblance$_2$) of the *logos* realized in the concrete token (260).

Here, the representational contents of the *pathema* qua *phantasma* and of the *pathema* qua meaning are nonidentical, but it is only the latter that is relevant to our understanding of the universal concept.

One might wonder why it is not the noetic act itself, the apprehension of the universal in the *phantasma*, which is the relevant *pathema*, rather than the *phantasma* itself. (This is not Modrak's position[3] although, again, we need to be careful that it does not seem to be her position. In saying that "meaning...is a way of thinking the content presented in the *phantasma*" (259), for example, it is not clear how meaning is any different from thinking, and thus that the relevant *pathema is* the *noema* (the thought).)

---

[2] How? By hearing or seeing the sign? But that is the association it is about to make.

[3] It is, however, the position advanced by Michael Wedin in *Mind and Imagination in Aristotle* (New Haven: Yale, 1988), chapter 3.

The above suggestion does not seem to be right, however. If the thinking act is the apprehending of the universal, it would seem that the linguistic act—that of meaning something or understanding something by a sign—is the associating of the universal with a sign (not associating the *thought* with a sign), and it is perhaps in virtue of such an association that we are able to apprehend the universal (that is, think it). This would explain how the mind can think the universal in the *phantasma*, without first thinking what it is going to think.

I am inclined to think that, for Aristotle, the relevant *pathema* is in fact the universal "imprinted in" the *phantasma*, rather than the whole *phantasma* in which the universal is imprinted. Human beings have *phantasmata* that are "bumpier" than those of other animals (that is, those capable of imagination). The universal coming to rest in the soul is a process of reinforcement, with the recurring features of various impressions being pounded deeper into us as the result of Aristotelian induction (*epagoge*). Our words, then, are signs of these impressions of various depths, salient in any *phantasma* that represents an individual belonging to an ontological hierarchy. We cannot think a universal without the mind focusing on it, and a word focuses the mind on an impression at one particular depth rather than at another.

Why, on Aristotle's account, should I apprehend the universal human in my *phantasma* of Socrates when I hear the word 'human', rather than apprehend the universal animal, which is also present in the *phantasma*? It is, I believe, because the word is associated with an impression of a certain depth, rather than the deeper impression nested within it. And this suggests that it is the universal imprinted in the *phantasma*, rather than the whole *phantasma*, that is the relevant *pathema*.

Modrak's purpose is primarily to articulate and explain Aristotle's theory of meaning, and to show it to be a more sophisticated theory than it is usually thought to be. In doing so she is not reconstructing Aristotle. By that I mean that she is not simply arguing that there is something like this theory of meaning available to Aristotle, and consistent with what he says (or at least with how we understand him). Modrak's general interpretation is implicit in what Aristotle says, it being the conception that fills in the details left unstated in *De Interpretatione* 1.

It is not Modrak's purpose, however, to justify and defend Aristotle's theory of meaning, at least not directly. But in presenting his theory as interrelated with his overall philosophy, and as having greater depth and subtlety than it is usually accorded, she simultaneously presents it as a viable and perhaps even arguably plausible theory.

The theory, thus more fully articulated, avoids a number of the difficulties that have been thought to confront Aristotle's theory when understood in terms of a more simplistic interpretation. Of course, this will not sway anyone who believes that the imagist program is hopelessly flawed (though it might lead them to reassess some of their reasons for thinking so), but that is not Modrak's present concern. (In the final chapter, she does give an, admittedly, all too brief survey of some outstanding issues from the modern perspective, but there is little here that will give the anti-imagist reason to think the Aristotelian account is viable.)

A final note is in order. The book advertises itself as "of particular interest to classical philosophers, classicists, and historians of psychology and cognitive science," and this is indeed appropriate. Its interest to classical philosophers, in particular, lies not only in its significant interpretive insight into Aristotle's account of the role of imagination in cognitive processes, but also in the path it traces through more familiar Aristotelian doctrine displaying a coherent and interconnected picture, viewed from the perspective of his theory of meaning. This is a path along which one needs to be led, in order to see the connections and have them explained.

The book's advertised audience, however, should be expanded to include contemporary philosophers of language (and mind), and philosophers more generally. For it provides an accessible and readable presentation of some of the more important components of Aristotelian doctrine, without going too deeply into the finer points of Aristotelian scholarship. Those who do want to delve more deeply are afforded ample references to the relevant literature, and helpful notes.

More importantly, it presents Aristotle's theory not as a primitive, naïve, and somewhat simplistic image theory of meaning, but as one of considerable substance with more than merely historical interest—a theory well worth rethinking in the context of contemporary discussions, both in the philosophy of language and in the philosophy of mind.

RICHARD TIERNEY

Western Michigan University

NEW BOOKS

ABOULAFIA, MITCHELL: *The Cosmopolitan Self: George Herbert Mead and Continental Philosophy*. Urbana: Illinois UP, 2001. x, 169 p. Cloth $29.95.

AGAR, NICHOLAS: *Life's Intrinsic Value: Science, Ethics, and Nature.* New York: Columbia, 2001. x, 200 p Cloth $45.00, paper $25.00.

AJZENSTAT, OONA: *Driven Back to the Text: The Premodern Sources of Levinas's Postmodernism.* Pittsburgh: Duquesne, 2001. ix, 388 p. Cloth $54.00.

ALEXANDER, LARRY, and EMILY SHERWIN: *The Rule of Rules: Morality, Rules, and the Dilemmas of Law.* Durham, NC: Duke, 2001. viii, 279 p. Cloth $45.95.

ALLEN, GEORGE: *The Patterns of the Present: Interpreting the Authority of Form.* Albany. SUNY, 2001. xvii, 305 p. Cloth $59.50, paper $19.95.

ALLEN, JAMES: *Inference from Signs: Ancient Debates about the Nature of Evidence.* New York: Oxford, 2001. x, 279 p. Cloth $45.00.

ALMOG, JOSEPH: *What Am I? Descartes and the Mind-Body Problem.* New York: Oxford, 2002. xviii, 139 p. Cloth $25.00.

ALTMAN, SIMON L.: *Is Nature Supernatural? A Philosophical Exploration of Science and Nature.* Amherst, NY: Prometheus, 2002. 680 p. Cloth $39.00.

AMES, ROGER T., and DAVID L. HALL: *Focusing the Familiar: A Translation and Philosophical Interpretation of the* Zhongyong. Honolulu: Hawaii UP, 2001. xiii, 165 p. Cloth $35.00, paper $17.95.

ARP, KRISTANA: *The Bonds of Freedom: Simone de Beauvoir's Existentialist Ethics.* Chicago: Open Court, 2001. xii, 178 p. Cloth $42.95, paper $19.95.

ASCHHEIM, STEVEN E.: *Scholem, Arendt, and Klemperer: Intimate Chronicles in Turbulent Times.* Bloomington: Indiana UP, 2001. 134 p. Cloth $27.95.

BALDWIN, THOMAS: *Contemporary Philosophy: Philosophy in English since 1945.* New York: Oxford, 2001. x, 294 p. Paper $18.95.

BALOT, RYAN K.: *Greed and Injustice in Classical Athens.* Princeton: University Press, 2001. x, 291 p. Cloth $39.50.

BARACCHI, CLAUDIA: *Of Myth, Life, and War in Plato's* Republic. Bloomington: Indiana UP, 2002. ix, 249 p. Cloth $49.95, paper $24.95.

BARBER, MICHAEL D.: *Equality and Diversity: Phenomenological Investigations of Prejudice and Discrimination.* Amherst, NY: Humanity, 2001. 296 p. Cloth $55.00.

BARNEY, RACHEL: *Names and Nature in Plato's* Cratylus. New York: Routledge, 2001. vii, 227 p. Cloth $70.00.

BATES, DAVID W.: *Enlightenment Aberrations: Error and Revolution in France.* Ithaca: Cornell, 2002. xiii, 262 p. Cloth $39.95.

BATTERMAN, ROBERT W.: *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence.* New York: Oxford, 2002. x, 144 p. Cloth $35.00.

BAUER, NANCY: *Simone de Beauvoir, Philosophy, and Feminism.* New York: Columbia, 2001. xii, 303 p. Cloth $49.50, paper $18.50.

BECK, JOHN: *Writing the Radical Center. William Carlos Williams, John Dewey, and American Cultural Politics.* Albany: SUNY, 2001. xii, 202 p Cloth $59.50, paper $19.95.

BELNAP, NUEL, MICHAEL PERLOFF, and MING XU: *Facing the Future: Agents and Choices in Our Indeterminist World.* New York: Oxford, 2001. xvi, 501 p. Cloth $49.95.

BENNETT, JONATHAN: *Learning from Six Philosophers: Descartes, Spinoza, Leibniz, Locke, Berkeley, Hume, Volume 1.* New York: Oxford, 2001. xix, 403 p. Cloth $45.00; *Learning from Six Philosophers: Descartes, Spinoza, Leibniz, Locke, Berkeley, Hume, Volume 2.* New York: Oxford, 2001. xix, 375 p. Cloth $45.00.

BENTON, TED, and IAN CRAIB: *Philosophy of Social Science: The Philosophical Foundations of Social Thought.* New York: Palgrave, 2001. vii, 203 p. Cloth $65.00, paper $19.95.

BENZON, WILLIAM: *Beethoven's Anvil: Music in Mind and Culture.* New York: Basic, 2001. xvi, 336 p. Cloth $27.50.

BERNSTEIN, J. M.: *Adorno: Disenchantment and Ethics.* New York: Cambridge, 2001. xviii, 460 p. Cloth $70.00, paper $25.00.

BEST, STEVEN, and DOUGLAS KELLNER: *The Postmodern Adventure: Science, Technology, and Cultural Studies at the Third Millennium.* New York: Guilford, 2001. 313 p. Cloth $38.00, paper $19.95.

BITTNER, RÜDIGER: *Doing Things for Reasons.* New York: Oxford, 2001. xi, 204 p. Cloth $35.00.

BLOCKER, H. GENE, and CHRISTOPHER L. STARLING: *Japanese Philosophy.* Albany: SUNY, 2001. 213 p. Cloth $62.50, paper $20.95.

BLOOMFIELD, PAUL: *Moral Reality.* New York: Oxford, 2001. xvi, 208 p. Cloth $39.95.

BLUM, LAWRENCE: *"I'm Not a Racist, But. . .": The Moral Quandary of Race.* Ithaca: Cornell, 2002. x, 259 p. Cloth $29.95.

BOOTHBY, RICHARD: *Freud as Philosopher: Metapsychology after Lacan.* New York: Routledge, 2001. xix, 330 p. Cloth $90.00, paper $24.95.

BOUDON, RAYMOND: *The Origin of Values: Sociology and Philosophy of Beliefs.* New Brunswick, NJ: Transaction, 2001. vii, 230 p. Cloth $39.95.

BRAYBROOKE, DAVID: *Natural Law Modernized.* Toronto: University Press, 2001. xiii, 351 p. Cloth $70.00.

BRISON, SUSAN J.: *Aftermath: Violence and the Remaking of a Self.* Princeton: University Press, 2002. xiii, 165 p. Cloth $19.95.

BROWN, JAMES ROBERT: *Who Rules in Science: An Opinionated Guide to the Wars.* Cambridge: Harvard, 2001. xiii, 236 p. Cloth $26.00.

BROWN, WENDY. *Politics out of History.* Princeton: University Press, 2001. x, 193 p. Cloth $45.00, paper $14.95.

BUNGE, MARIO: *Philosophy in Crisis: The Need for Reconstruction.* Amherst, NY: Prometheus, 2001. 245 p. Cloth $33.00.

CAHOONE, LAWRENCE E.: *Civil Society. The Conservative Meaning of Liberal Politics.* Malden, MA: Blackwell, 2002. vi, 316 p. Paper $27.95.

CAMP, JOSEPH L., JR : *Confusion. A Study in the Theory of Knowledge.* Cambridge: Harvard, 2002. x, 246 p. Cloth $39.95.

CANNON, BOB: *Rethinking the Normative Content of Critical Theory: Marx, Habermas, and Beyond.* New York: Palgrave, 2001. xiv, 211 p. Cloth $65.00.

CARROLL, NOËL: *Beyond Aesthetics: Philosophical Essays.* New York: Cambridge, 2001. xiv, 450 p. Cloth $74.95.

CARTER, ROBERT E.: *Encounter with Enlightenment: A Study of Japanese Ethics.* Albany: SUNY, 2001. xxxiii, 258 p. Paper $22.95.

CHAKRAVARTI, SITANSU S.: *Modality, Reference, and Sense: An Essay in the Philosophy of Language.* New Delhi: Munshiram Manoharlal, 2001. xii, 148 p. Cloth $27.50.

CHANTER, TINA: *Time, Death, and the Feminine. Levinas with Heidegger.* Stanford: University Press, 2001. xix, 294 p. Cloth $55.00, paper $21.95.

CLIFFORD, MICHAEL: *Political Genealogy after Foucault: Savage Identities.* New York: Routledge, 2001. xiii, 225 p. Paper $24.95.

COAKLEY, SARAH: *Powers and Submissions: Spirituality, Philosophy and Gender.* Malden, MA: Blackwell, 2002. xx, 172 p. Cloth $62.95.

COCKBURN, DAVID: *An Introduction to the Philosophy of Mind.* New York: Palgrave, 2001 xviii, 157 p. Cloth $55.00, paper $18.85.

COHEN, RICHARD A.: *Ethics, Exegesis and Philosophy: Interpretation after Levinas.* New York: Cambridge, 2001. vi, 361 p. Cloth $64.95.

COLAIACO, JAMES A.: *Socrates against Athens: Philosophy on Trial.* New York: Routledge, 2001. xi, 266 p. Cloth $85.00, paper $22.95.

COLEMAN, JULES: *The Practice of Principle: In Defence of a Pragmatist Approach to Legal Theory.* New York: Oxford, 2001. xx, 226 p. Cloth $39.95.

CONSIGNY, SCOTT: *Gorgias: Sophist and Artist.* Columbia: South Carolina UP, 2001. xiii, 242 p. Cloth $39 95.

CORDNER, CHRISTOPHER: *Ethical Encounter: The Depth of Moral Meaning.* New York: Palgrave, 2002. viii, 208 p. Cloth $62.00.

COXON, A. H.: *The Philosophy of Form: An Analytical and Historical Commentary on Plato's* Parmenides. Assen, The Netherlands: Van Gorcum, 1999. 172 p. Cloth $32.50.

CRAIG, LEON HAROLD: *Of Philosophers and Kings: Political Philosophy in Shakespeare's* Macbeth *and* King Lear. Toronto: University Press, 2001. xi, 406 p. Cloth $70.00.

CRAIG, WILLIAM LANE: *God, Time, and Eternity: The Coherence of Theism II: Eternity.* Norwell, MA: Kluwer, 2001. xi, 321 p. Cloth $105.00.

CUMMING, ROBERT DENOON: *Phenomenology and Deconstruction: Breakdown in Communication, Volume 3.* Chicago: University Press, 2001. x, 316 p. Cloth $56.00, paper $27.00; *Phenomenology and Deconstruction: Solitude, Volume 4.* Chicago: University Press, 2001. xi, 222 p. Cloth $50.00, paper $22.00.

CUNNINGHAM, ANTHONY: *The Heart of What Matters: The Role for Literature in Moral Philosophy.* Berkeley: California UP, 2001. x, 296 p. Cloth $60.00, paper $24.95.

DAINTON, BARRY. *Time and Space.* Montreal: McGill-Queen, 2001. xiv, 386 p. Cloth $75.00, paper $29.95.

DASGUPTA, PARTHA: *Human Well-Being and the Natural Environment.* New York: Oxford, 2001. xxii, 305 p. Cloth $35.00.

DAVIDSON, ARNOLD I.: *The Emergence of Sexuality: Historical Epistemology and the Formation of Concepts.* Cambridge: Harvard, 2001. xvi, 254 p. Cloth $39.95.

DAVIDSON, DONALD: *Subjective, Intersubjective, Objective.* New York: Oxford, 2001. xviii, 237 p. Paper $19.95.

DAVIE, GEORGE: *The Scotch Metaphysics: A Century of Enlightenment in Scotland.* New York: Routledge, 2001. xi, 241 p. Cloth $100.00.

DAVIES, RICHARD: *Descartes: Belief, Scepticism and Virtue.* New York: Routledge, 2001. xi, 371 p. Cloth $90.00.

DAVIES, STEPHEN: *Musical Works and Performances: A Philosophical Exploration.* New York: Oxford, 2001. viii, 374 p. Cloth $49.95.

DE VRIES, HENT: *Religion and Violence: Philosophical Perspectives from Kant to Derrida.* Baltimore: Johns Hopkins, 2002. xxiii, 443 p. Cloth $55.00, paper $24.95.

DEELY, JOHN: *Four Ages of Understanding: The First Postmodern Survey of Philosophy from Ancient Times to the Turn of the Twenty-First Century.* Toronto: University Press, 2001. viii, 1019 p. Cloth $95.00; *What Distinguishes Human Understanding?* South Bend, IN: St. Augustine, 2002. xiv, 178 p. Cloth $25.00.

DELANCEY, CRAIG: *Passionate Engines: What Emotions Reveal about Mind and Artificial Intelligence.* New York: Oxford, 2002. xvi, 254 p. Cloth $49.95.

DEMBSKI, WILLIAM A.: *No Free Lunch: Why Specified Complexity Cannot Be Purchased without Intelligence.* Lanham, MD: Rowman and Littlefield, 2002. xxv, 404 p. Cloth $35.00.

DEWEIL, BORIS: *Democracy: A History of Ideas.* Seattle: Washington UP, 2001. vi, 202 p. Paper $22.50.

DICKIE, GEORGE: *Art and Value.* Malden, MA, 2001. ix, 241 p. Cloth $56.95, paper $21.95.

DIGESER, P. E.: *Political Forgiveness.* Ithaca: Cornell, 2001. ix, 224 p. Cloth $39.95.

DOIG, JAMES C.: *Aquinas's Philosophical Commentary on the* Ethics. Norwell, MA: Kluwer, 2001. xvii, 311 p. Cloth $105.00.

DOOLEY, MARK: *The Politics of Exodus: Kierkegaard's Ethics of Responsibility.* New York: Fordham, 2001. xxiv, 285 p. Cloth $40.00, paper $20.00.

DREYFUS, HUBERT L.: *On the Internet.* New York: Routledge, 2001. ix, 127 p. Cloth $50.00, paper $12.95.

DUMMETT, MICHAEL: *On Immigration and Refugees.* New York: Routledge, 2001. xiii, 160 p. Cloth $50.00, paper $12.95.

DUPRÉ, JOHN: *Human Nature and the Limits of Science.* New York: Oxford, 2001. x, 201 p. Cloth $24.95.

DURAN, JANE: *Worlds of Knowing: Global Feminist Epistemologies.* New York: Routledge, 2001. xvi, 304 p. Cloth $85.00, paper $23.95.

EDEL, ABRAHAM: *Ethical Theory and Social Change: The Evolution of John Dewey's Ethics, 1908-1932.* New Brunswick, NJ: Transaction, 2001. xxii, 215 p. Cloth $39.95.

ELLIS, BRIAN: *Scientific Essentialism.* New York: Cambridge, 2001. xiv, 309 p. Cloth $54.95.

ESFELD, MICHAEL: *Holism in Philosophy of Mind and Philosophy of Physics.* Norwell, MA: Kluwer, 2001. xiv, 366 p. Cloth $113.00.

EZE, EMMANUEL CHUKWUDI: *Achieving Our Humanity: The Idea of the Postracial Future.* New York: Routledge, 2001. xviii, 250 p. Cloth $85.00, paper $23.95.

FENN, RICHARD K.: *Beyond Idols: The Shape of a Secular Society.* New York: Oxford, 2001. 196 p. Cloth $27.50.

FENVES, PETER: *Arresting Language: From Leibniz to Benjamin.* Stanford: University Press, 2001. xiii, 379 p. Cloth $55.00, paper $24.95.

FERREIRA, M. JAMIE: *Love's Grateful Striving: A Commentary on Kierkegaard's* Works of Love. New York: Oxford, 2001. xi, 316 p. Cloth $45.00.

FIELD, HARTRY: *Truth and the Absence of Fact.* New York: Oxford, 2001. xv, 401 p. Cloth $24.95.

FLORKA, ROGER: *Descartes's Metaphysical Reasoning.* New York: Routledge, 2001. xii, 124 p. Cloth $60.00.

FOGELIN, ROBERT: *Berkeley and the Principles of Human Knowledge.* New York: Routledge, 2001. xvi, 166 p. Cloth $65.00, paper $14.95.

FOLEY, RICHARD: *Intellectual Trust in Oneself and Others.* New York: Cambridge, 2001. x, 182 p. Cloth $55.00.

FOOT, PHILIPPA: *Natural Goodness.* New York: Oxford, 2001. 125 p. Cloth $22.00.

FORRAI, GÁBOR: *Reference, Truth, and Conceptual Schemes: A Defense of Internal Realism.* Norwell, MA: Kluwer, 2001. 150 p. Cloth $49.00.

FOSS, JEFFREY: *Science and the Riddle of Consciousness: A Solution.* Norwell, MA: Kluwer, 2000. xiii, 225 p. Cloth $99.50.

FRIEDMAN, MICHAEL: *Dynamics of Reason.* Stanford: CSLI, 2001. xiv, 141 p. Cloth $55.00, paper $19.50.

FROST, LAURA: *Sex Drives: Fantasies of Fascism in Literary Modernism.* Ithaca: Cornell, 2002. viii, 197 p. Cloth $45.00, paper $17.95.

GANERI, JONARDON: *Philosophy in Classical India.* New York: Routledge, 2001. xi, 207 p. Paper $24.95.

GARFIELD, JAY L.: *Empty Words: Buddhist Philosophy and Cross-cultural Interpretations.* New York: Oxford, 2002. xiv, 306 p. Paper $24.95.

GEORGE, ALEXANDER, and DANIEL J. VELLEMAN: *Philosophies of Mathematics.* Malden, MA: Blackwell, 2002. viii, 230 p. Paper $29.95.

GEUSS, RAYMOND: *Public Goods, Private Goods.* Princeton: University Press, 2001. viii, 148 p. Cloth $19.95.

GILLROY, JOHN MARTIN: *Justice and Nature: Kantian Philosophy, Environmental Policy, and the Law.* Washington, DC: Georgetown, 2001. xxxix, 444 p. Cloth $70.00.

GLANNON, WALTER: *Genes and Future People: Philosophical Issues in Human Genetics.* Boulder, CO: Westview, 2001. x, 216 p. Cloth $80.00, paper $26.00.

GOLDBERG, DAVID THEO: *The Racial State.* Malden, MA: Blackwell, 2002. x, 319 p. Paper $24.95.

GOLDMAN, ALAN H.: *Practical Rules: When We Need Them and When We Don't.* New York: Cambridge, 2002. xi, 210 p. Cloth $54.95.

GOLDMAN, ALVIN I.: *Pathways to Knowledge: Private and Public.* New York: Oxford, 2002. xiv, 224 p. Cloth $35.00.

GOODING-WILLIAMS, ROBERT: *Zarathustra's Dionysian Modernism.* Stanford: University Press, 2001. x, 420 p. Cloth $55.00, paper $24.95.

GOODMAN, LENN E.: *In Defense of Truth: A Pluralistic Approach.* Amherst, NY: Prometheus, 2001. 431 p. Cloth $55.00.

GORDON, HAIM: *The Heidegger-Buber Controversy: The Status of the I-Thou.* Westport, CT: Greenwood, 2001. xix, 170 p. Cloth $66.00.

GÖTZ, IGNACIO L.: *Technology and the Spirit.* Westport, CT: Praeger, 2001. xv, 137 p. Cloth $53.00.

GRAFTON, ANTHONY: *Bring Out Your Dead: The Past as Revelation.* Cambridge: Harvard, 2001. vii, 360 p. Cloth $39.95.

GRANT, EDWARD: *God and Reason in the Middle Ages.* New York: Cambridge, 2001. ix, 397 p. Cloth $64.95.

GREEN, KAREN: *Dummett: Philosophy of Language.* Cambridge, United Kingdom: Polity, 2001. xi, 236 p. Cloth $62.95, paper $24.95.

GREGOR, JAMES A.: *Giovanni Gentile: Philosopher of Fascism.* New Brunswick, NJ: Transaction, 2001. xiv, 125 p. Cloth $29.95.

GRIFFITHS, PAUL J.: *Problems of Religious Diversity.* Malden, MA: Blackwell, 2001. xv, 176 p. Cloth $59.50, paper $24.95.

HACKER, P. M. S.: *Wittgenstein: Connections and Controversies.* New York: Oxford, 2001. xxiv, 375 p. Cloth $49.95.

HACKING, IAN: *An Introduction to Probability and Inductive Logic.* New York: Cambridge, 2001. xvii, 301 p. Cloth $69.95, paper $24.95.

HALE, BOB, and CRISPIN WRIGHT: *The Reason's Proper Study: Essays Towards a Neo-Fregean Philosophy of Mathematics.* New York: Oxford, 2001. xiv, 455 p. Cloth $55.00.

HANDS, D. WADE: *Reflection without Rules: Economic Methodology and Contemporary Science Theory.* New York: Cambridge, 2001. xi, 480 p. Cloth $95.00, paper $34.95.

HANNAY, ALASTAIR: *Kierkegaard: A Biography.* New York: Cambridge, 2001. xvi, 496 p. Cloth $39.95.

HANSSON, SVEN OVE: *The Structure of Values and Norms.* New York: Cambridge, 2001. xiii, 314 p. Cloth $59.95.

HARE, JOHN E.: *God's Call: Moral Realism, God's Commands, and Human Autonomy.* Grand Rapids, MI: Eerdmans, 2001. x, 122 p. Cloth $14.00.

HARNISH, ROBERT M.: *Minds, Brains, and Computers: An Historical Introduction to the Foundations of Cognitive Science.* Malden, MA: Blackwell, 2002. xiv, 446 p. Cloth $64.95, paper $32.95.

HAY, PETER: *Main Currents in Western Environmental Thought.* Bloomington: Indiana UP, 2002. x, 400 p. Cloth $59.95, paper $29.95.

HAYES, CALVIN: *Fallibilism Democracy and the Market: The Meta-theoretical Foundations of Popper's Political Philosophy.* Lanham, MD: UP of America, 2001. xxi, 223 p. Cloth $35.00.

HELM, BENNETT W.: *Emotional Reason: Deliberation, Motivation, and the Nature of Value.* New York: Cambridge, 2001. x, 261 p. Cloth $59.95.

HESK, JON: *Deception and Democracy in Classical Athens.* New York: Cambridge, 2000. viii, 336 p. Cloth $64.95.

HORNER, ROBYN: *Rethinking God as Gift: Marion, Derrida, and the Limits of Phenomenology.* New York: Fordham, 2001. xi, 272 p. Cloth $40.00, paper $20.00.

HOROWITZ, GREGG M.: *Sustaining Loss: Art and Mournful Life.* Stanford: University Press, 2001. xiv, 239 p. Cloth $49.50, paper $19.95.

HUDSON, HUD: *A Materialist Metaphysics of the Human Person.* Ithaca: Cornell, 2001. xii, 202 p. Cloth $39.95.

HUEMER, MICHAEL: *Skepticism and the Veil of Perception.* Lanham, MD: Rowman and Littlefield, 2001. xx, 209 p. Cloth $60.00, paper $23.95.

HUGHES, GERARD J.: *Aristotle on Ethics.* New York: Routledge, 2001. x, 238 p. Cloth $60.00, paper $14.95.

HUTCHINSON, BRIAN: *G.E. Moore's Ethical Theory: Resistance and Reconciliation.* New York: Cambridge, 2001. viii, 219 p. Cloth $54.95.

ISRAEL, JONATHAN I.: *Radical Enlightenment: Philosophy and the Making of Modernity, 1650-1750.* New York: Oxford, 2001. xix, 810 p. Cloth $45.00.

JACOBS, ALAN: *A Theology of Reading: The Hermeneutics of Love.* Boulder, CO: Westview, 2001. ix, 186 p. Cloth $65.00, paper $18.00.

JHA, STEFANIA RUZSITS: *Reconsidering Michael Polanyi's Philosophy.* Pittsburgh: University Press, 2002. vii, 318 p. Cloth $32.00.

KAHN, CHARLES H.: *Pythagoras and the Pythagoreans: A Brief History.* Indianapolis: Hackett, 2001. xi, 195 p. Cloth $34.95, paper $14.95.

KAVANAUGH, JOHN F.: *Who Count as Persons? Human Identity and the Ethics of Killing.* Washington, DC: Georgetown, 2001. xiv, 233 p. Paper $24.95.

KEARNEY, RICHARD: *The God Who May Be: A Hermeneutics of Religion.* Bloomington: Indiana UP, 2001. ix, 172 p. Cloth $49.95, paper $19.95; *On Stories: Thinking in Action.* New York: Routledge, 2002. xi, 193 p. Paper $12.95.

KELLENBERGER, J.: *Moral Relativism, Moral Diversity and Human Relationships.* University Park: Penn State UP, 2001. ix, 236 p. Cloth $35.00.

KITCHER, PHILIP: *Science, Truth, and Democracy.* New York: Oxford, 2001. 219 p. Cloth $27.50.

KIVY, PETER: *New Essays on Musical Understanding.* New York: Oxford, 2001. x, 236 p. Paper $19.95; *The Possessor and the Possessed: Handel, Mozart, Beethoven, and the Idea of Musical Genius.* New Haven: Yale, 2001. xiv, 287 p. Cloth $35.00.

KOLTUN-FROMM, KEN: *Moses Hess and Modern Jewish Identity.* Bloomington: Indiana UP, 2001. x, 180 p. Cloth $32.95.

KOSKY, JEFFREY L.: *Levinas and the Philosophy of Religion.* Bloomington: Indiana UP, 2001. xxiv, 233 p. Cloth $39.95.

KOSSO, PETER: *Knowing the Past: Philosophical Issues of History and Archaeology.* Amherst, NY: Humanity, 2001. 200 p. Cloth $45.00.

KRANTZ, SUSAN LUFKIN: *Refuting Peter Singer's Theory: The Importance of Human Dignity.* Westport, CT: Praeger, 2002. xvi, 133 p. Cloth $54.95.

KRISTJÁNSSON, KRISTJÁN: *Justifying Emotions: Pride and Jealousy.* New York: Routledge, 2002. xii, 257 p. Cloth $80.00.

KUKLA, ANDRÉ: *Methods of Theoretical Psychology.* Cambridge: MIT, 2001. xii, 250 p. Cloth $40.00.

KUMAR, RAHUL: *Consensualism in Principle: On the Foundations of Non-consequentialist Moral Reasoning.* New York: Routledge, 2001. xii, 166 p. Cloth $60.00.

KURTZ, PAUL: *Skepticism and Humanism: The New Paradigm.* New Brunswick, NJ: Transaction, 2001. 306 p. Cloth $39.95

KWINTER, SANFORD: *Architectures of Time: Toward a Theory of the Event in Modernist Culture.* Cambridge: MIT, 2001. x, 237 p. Cloth $29.95.

KYBURG, HENRY E., JR., and CHOH MAN TENG: *Uncertain Inference.* New York: Cambridge, 2001. xii, 298 p. Cloth $90.00, paper $32.95.

KYMLICKA, WILL: *Politics in the Vernacular: Nationalism, Multiculturalism, and Citizenship.* New York: Oxford, 2001. 383 p. Paper $18.95.

LA CAZE, MARGUERITE: *The Analytic Imaginary.* Ithaca: Cornell, 2002. ix, 194 p. Cloth $32.50.

LA VOPA, ANTHONY J.: *Fichte: The Self and the Calling of Philosophy, 1762-1799.* New York: Cambridge, 2001. xiv, 449 p. Cloth $54.95.

LACEY, A. R.: *Robert Nozick.* Princeton: University Press, 2001. vii, 248 p. Cloth $55.00, paper $17.95.

LAMPERT, LAURENCE: *Nietzsche's Task: An Interpretation of* Beyond Good and Evil. New Haven: Yale, 2001. x, 320 p. Cloth $40.00.

LAWLOR, KRISTA: *New Thoughts about Old Things: Cognitive Policies as the Ground of Singular Concepts.* New York: Garland, 2001. xiv, 122 p. Cloth $60.00.

LAWSON, HILARY: *Closure: A Story of Everything.* New York: Routledge, 2001. xlviii, 372 p. Cloth $90.00, paper $24.95.

LERNER, BEREL DOV: *Rules, Magic, and Instrumental Reason.* New York: Routledge, 2002. viii, 179 p. Cloth $90.00.

LESLIE, JOHN: *Infinite Minds: A Philosophical Cosmology.* New York: Oxford, 2001. x, 234 p. Cloth $35.00.

LEVINE, ANDREW: *Engaging Political Philosophy: From Hobbes to Rawls.* Malden, MA: Blackwell, 2002. vii, 274 p. Cloth $59.95, paper $24.95.

LILLA, MARK: *The Reckless Mind: Intellectuals in Politics.* New York: NYRB, 2001. xiii, 216 p. Cloth $24.95.

LIN, MIN: *Certainty as a Social Metaphor: The Social and Historical Production of Certainty in China and the West.* Westport, CT: Greenwood, 2001. xiv, 286 p. Cloth $72.50.

LLEWELYN, JOHN: *Appositions of Jacques Derrida and Emmanuel Levinas.* Bloomington: Indiana UP, 2002. xx, 255 p. Cloth $49.95, paper $24.95.

LONERGAN, BERNARD: *Phenomenology and Logic: The Boston College Lectures on Mathematical Logic and Existentialism.* PHILIP J. MCSHANE, ed. Toronto: University Press, 2001. xxiv, 419 p. Cloth $75.00, paper $26.95.

LONG, PETER: *Logic, Form and Grammar.* New York: Routledge, 2001. x, 110 p. Cloth $65.00.

LONGINO, HELEN E.: *The Fate of Knowledge.* Princeton: University Press, 2002. ix, 233 p. Cloth $49.50, paper $16.95.

LORENTZEN, JAMIE: *Kierkegaard's Metaphors.* Macon GA: Mercer, 2001. xiii, 201 p. $39.95.

LOVIBOND, SABINA: *Ethical Formation.* Cambridge: Harvard, 2002. xii, 203 p. Cloth $39.95.

LYCAN, WILLIAM G.: *Real Conditionals.* New York: Oxford, 2001. 223 p. Cloth $35.00.

LYONS, WILLIAM: *Matters of the Mind.* New York: Routledge, 2001. xxix, 288 p. Cloth $85.00, paper $22.95.

MADISON, GARY BRENT: *The Politics of Postmodernity: Essays in Applied Hermeneutics.* Norwell, MA: Kluwer, 2001. xv, 273 p. Cloth $99.00.

MAGEE, BRYAN: *The Tristan Chord: Wagner and Philosophy.* New York: Metropolitan, 2000. xi, 397 p. Cloth $35.00.

MAHDI, MUHSIN S.: *Alfarabi and the Foundation of Islamic Political Philosophy.* Chicago: University Press, 2001. xvii, 264 p. Cloth $37.50.

MAKIN, GIDEON: *The Metaphysics of Meaning: Russell and Frege on Sense and Denotation.* New York: Routledge, 2000. viii, 229 p. Paper $29.95.

MARGOLIS, JOSEPH: *Selves and Other Texts: The Case for Cultural Realism.* University Park: Penn State UP, 2001. xiv, 208 p. Cloth $29.95.

MASLIN, K. T.: *An Introduction to the Philosophy of Mind.* Malden, MA: Blackwell, 2001. xviii, 332 p. Cloth $59.95, paper $27.95.

MATHEWES, CHARLES T.: *Evil and the Augustinian Tradition.* New York: Cambridge, 2001. xii, 271 p. Cloth $59.95.

MATUŠTÍK, MARTIN BECK: *Jürgen Habermas: A Philosophical-Political Profile.* Lanham, MD: Rowman and Littlefield, 2001. xxxvii, 341 p. Cloth $85.00, paper $29.95.

MAZZOTTA, GIUSEPPE: *Cosmopoiesis: The Renaissance Experiment.* Toronto: University Press, 2001. xvi, 106 p. Cloth $35.00, paper $16.95.

MCBRIDE, WILLIAM L.: *From Yugoslav Praxis to Global Pathos: Anti-Hegemonic Post-post-marxist Essays.* Lanham, MD: Rowman and Littlefield, 2001. xvi, 249 p. Cloth $70.00, paper $24.95.

MCCARTHY, GEORGE E.: *Objectivity and the Silence of Reason: Weber, Habermas, and the Methodological Disputes in German Society.* New Brunswick, NJ: Transaction, 2001. 341 p. Cloth $54.95.

MCCOMISKEY, BRUCE: *Gorgias and the New Sophistic Rhetoric.* Carbondale: Southern Illinois UP, 2002. 156 p. Cloth $35.00.

MCDERMOTT, DREW V.: *Mind and Mechanism.* Cambridge: MIT, 2001. xiii, 262 p. Cloth $32.95.

MCMAHAN, JEFF: *The Ethics of Killing: Problems at the Margins of Life.* New York: Oxford, 2002. xiii, 540 p. Cloth $39.95.

MCMAHON, CHRISTOPHER: *Collective Rationality and Collective Reasoning.* New York: Cambridge, 2001. ix, 251 p. Cloth $54.95, paper $19.95.

MCSHERRY, CORYNNE: *Who Owns Academic Work? Battling for Control of Intellectual Property.* Cambridge: Harvard 2001. ix, 275 p. Cloth $29.95.

MEDEARIS, JOHN: *Joseph Schumpeter's Two Theories of Democracy.* Cambridge: Harvard, 2001. viii, 263 p. Cloth $45.00.

MELE, ALFRED R.: *Autonomous Agents: From Self Control to Autonomy.* New York: Oxford, 2001. viii, 271 p. Paper $19.95.

MERRICKS, TRENTON: *Object and Persons.* New York: Oxford, 2001. xii, 203 p. Cloth $35.00.

MEYERS, DIANA TIETJENS: *Gender in the Mirror: Cultural Imagery and Women's Agency.* New York: Oxford, 2002. 232 p. Paper $19.95.

MILLER, BARRY: *The Fullness of Being.* Notre Dame, IN: University Press, 2002. x, 175 p. Cloth $29.95.

MOELLENDORF, DARREL: *Cosmopolitan Justice.* Boulder, CO: Westview, 2002. xiii, 226 p. Cloth $75.00, paper $25.00.

MOORE, ADAM D.: *Intellectual Property and Information Control: Philosophic Foundations and Contemporary Issues.* New Brunswick, NJ: Transaction, 2001. xiv, 252 p. Cloth $34.95.

MOORE, GREGORY: *Nietzsche, Biology and Metaphor.* New York: Cambridge, 2002 viii, 228 p. Cloth $55.00.

MOORE, MARGARET: *The Ethics of Nationalism.* New York: Oxford, 2001. vii, 260 p. Cloth $45.00.

MORELAND, J. P.: *Universals.* Ithaca, NY: McGill-Queen's, 2001. viii, 184 p. Cloth $65.00, paper $27.95.

MORGAN, MICHAEL L.: *Interim Judaism: Jewish Thought in a Century of Crisis.* Bloomington: Indiana UP, 2001. xv, 146 p. Cloth $37.95, paper $15.95.

MULHALL, STEPHEN: *Inheritance and Originality: Wittgenstein, Heidegger, Kierkegaard.* New York: Oxford, 2001. xi, 448p. Cloth $60.00.

MURPHY, MARK C.: *Natural Law and the Practical Rationality.* New York: Cambridge, 2001. xiv, 284 p. Cloth $54.95.

NARDIN, TERRY: *The Philosophy of Michael Oakeshott.* University Park: Penn State UP, 2001. x, 241 p. Cloth $35.00.

NAVIA, LUIS E.: *Antisthenes of Athens: Setting the World Aright.* Westport, CT: Greenwood, 2001. xii, 176 p. Cloth $62.00.

NEWMAN, JAY: *Biblical Religion and Family Values: A Problem in the Philosophy of Culture.* Westport, CT: Praeger, 2001. 357 p. Cloth $79.95.

NICHOLSEN, SHIERRY WEBER: *The Love of Nature and the End of the World: The Unspoken Dimensions of Environmental Concern.* Cambridge: MIT, 2002. x, 216 p. Cloth $27.95.

NIELSEN, KAI: *Naturalism and Religion.* Amherst, NY: Prometheus, 2001. 506 p. Cloth $40.00.

NOZICK, ROBERT: *Invariances: The Structure of the Objective World.* Cambridge: Harvard, 2001. x, 416 p. Cloth $35.00.

NUCCETELLI, SUSANA: *Latin American Thought: Philosophical Problems and Arguments.* Boulder, CO: Westview, 2002. xxii, 268 p. Cloth $75.00, paper $25.00.

NUSSBAUM, MARTHA C.: *Upheavals of Thought: The Intelligence of Emotions.* New York: Cambridge, 2001. xii, 722 p. Cloth $39.95.

OBERDIEK, HANS: *Tolerance: Between Forbearance and Acceptance.* Lanham, MD: Rowman and Littlefield, 2001. ix, 181 p. Cloth $70.00, paper $24.95.

O'CONNOR, DAVID: *Hume on Religion.* New York: Routledge, 2001. xv, 227 p. Cloth $60.00, paper $14.95.

OLAFSON, FREDERICK A.: *Naturalism and the Human Condition: Against Scientism.* New York: Routledge, 2001. xi, 115 p. Cloth $75.00, paper $22.95.

OSTROW, MATTHEW B.: *Wittgenstein's* Tractatus: *A Dialectical Interpretation.* New York: Cambridge, 2002. xi, 175 p. Cloth $55.00, paper $20.00.

PARSONS, SUSAN FRANK: *The Ethics of Gender.* Malden, MA: Blackwell, 2002. x, 202 p. Paper $24.95.

PASNAU, ROBERT: *Thomas Aquinas on Human Nature: A Philosophical Study of* Summa Theologiae *1a 75-89.* New York: Cambridge, 2002. xi, 500 p. Cloth $75.00, paper $28.00.

PEACH, LUCINDA: *Legislating Morality: Pluralism and Religious Identity in Lawmaking.* New York: Oxford, 2002. viii, 218 p. Cloth $35.00.

PEARSON, KEITH ANSELL: *Philosophy and the Adventure of the Virtual: Bergson and the Time of Life.* New York: Routledge, 2002. ix, 246 p. Cloth $80.00, paper $25.95.

PEREGRIN, JAROSLAV: *Meaning and Structure: Structuralism of (Post) Analytic Philosophers.* Burlington, VT: Ashgate, 2001. vii, 276 p. Cloth $69.95.

PESIC, PETER: *Seeing Double: Shared Identities in Physics, Philosophy, and Literature.* Cambridge: MIT, 2002. 184 p. Cloth $24.95.

PETTIT, PHILIP: *A Theory of Freedom.* New York: Oxford, 2001. 193 p. Paper $19.95.

PHILLIPS, D. Z.: *Religion and the Hermeneutics of Contemplation.* New York: Cambridge, 2001. xiv, 330 p. Cloth $64.95, paper $24.95.

PINES, YURI: *Foundations of Confucian Thought: Intellectual Life in the Chunqiu Period, 722-453 B.C.E.*. Hawaii: University Press, 2002. ix, 387 p. Cloth $55.00.

PRICE, CAROLYN: *Functions in Mind: A Theory of Intentional Content.* New York: Oxford, 2001. vi, 263 p. Cloth $49.95.

PRIEST, GRAHAM: *An Introduction to Non-Classical Logic.* New York: Cambridge, 2001. xxi, 242 p. Cloth $54.95, paper $19.95.

PUTNAM, HILARY: *Enlightenment and Pragmatism.* Assen, The Netherlands: Van Gorcum, 2001. 48 p. Paper.

RAPHAEL, D. D.: *Concepts of Justice.* New York: Oxford, 2001. 256 p. Cloth $35.00.

RAZ, JOSEPH: *Value, Respect, and Attachment.* New York: Cambridge, 2001. vii, 178 p. Cloth $54.95, paper $18.95.

REEVE, C. D. C.: *Women in the Academy: Dialogues on Themes from Plato's* Republic. Indianapolis: Hackett, 2001. viii, 72 p. Cloth $27.95, paper $6.95.

RESCHER, NICHOLAS: *Cognitive Pragmatism: The Theory of Knowledge in Pragmatic Perspective.* Pittsburgh: University Press, 2001. xi, 250 p. Cloth $27.50; *Minding Matter and Other Essays in Philosophical Inquiry.* Lanham, MD: Rowman and Littlefield, 2001. xii, 147 p. Cloth $60.00, paper $21.95; *Nature and Understanding: The Metaphysics and Method of Science.* New York: Oxford, 2000. x, 186 p. Cloth $35.00; *Paradoxes: Their Roots, Range, and Resolution.* Chicago: Open Court, 2001. xxiii, 293 p. Cloth $52.95, paper $24.95; *Philosophical Reasoning: A Study in the Methodology of Philosophizing.* Malden, MA: Blackwell, 2001. xi, 282 p. Cloth $69.95, paper $34.95.

RICHARDS, GLYN: *Ghandhi's Philosophy of Education.* New York: Oxford, 2001. viii, 118 p. Cloth $15.95.

ROSENBERG, ALEX: *Philosophy of Science: A Contemporary Introduction.* New York: Routledge, 2000. ix, 191 p. Cloth $80.00, paper $27.95.

ROWLANDS, MARK: *The Nature of Consciousness.* New York: Cambridge, 2001. ix, 245 p. Cloth $60.00.

ROY, LOUIS: *Transcendent Experiences: Phenomenology and Critique.* Toronto: University Press, 2001. xiv, 219 p. Cloth $60.00.

SANTAS, GERASIMOS: *Goodness and Justice: Plato, Aristotle, and the Moderns.* Malden, MA: Blackwell, 2001. x, 300 p. Cloth $64.95, paper $29.95.

SAVAGE, DANIEL M.: *John Dewey's Liberalism: Individual, Community, and Self-development.* Carbondale: Southern Illinois UP, 2002. ix, 219 p. Cloth $40.00.

SAVILE, ANTHONY: *Leibniz and the* Monadology. New York: Routledge, 2000. v, 247 p. Cloth $60.00, paper $14.99.

SAXENA, S. K.: *Hindustani Sangeet and a Philosopher of Art: Music, Rhythm and Kathak Dance vis-à-vis Aesthetics of Susanne K. Langer.* New Delhi: D.K. Printworld, 2001. x, 383 p. Cloth $30.00.

SCHAIN, RICHARD: *The Legend of Nietzsche's Syphilis.* Westport, CT: Greenwood, 2001. xiii, 130 p. Cloth $62.95.

SCHEFFLER, SAMUEL: *Boundaries and Alleigances: Problems of Justice and Responsibility in Liberal Thought.* New York: Oxford, 2001. 221 p. Cloth $29.95.

SCHMIDT, DENNIS J.: *On Germans and Other Greeks: Tragedy and Ethical Life.* Bloomington: Indiana UP, 2001. xviii, 337 p. Cloth $49.95, paper $24.95.

SCHULTZ, WALTER J.: *The Moral Conditions of Economic Efficiency.* New York: Cambridge, 2001. xii, 144 p. Cloth $49.95.

SEARLE, JOHN R.: *Rationality in Action.* Cambridge: MIT, 2001. xvi, 303 p. Cloth $35.00.

SIDER, THEODORE: *Four-Dimensionalism: An Ontology of Persistence and Time.* New York: Oxford, 2001. xxiv, 255 p. Cloth $45.00.

SIMPSON, LORENZO C.: *The Unfinished Project: Toward a Postmetaphysical Humanism.* New York: Routledge, 2001. xi, 180 p. Cloth $85.00, paper $22.95.

SINGER, IRVING: *Explorations in Love and Sex.* Lanham, MD: Rowman and Littlefield, 2001. xiii, 239 p. Cloth $29.95; *Feelings and Imagination: The Vibrant Flux of Our Existence.* Lanham, MD: Rowman and Littlefield, 2001. xii, 223 p. Cloth $35.00; *Sex: A Philosophical Primer.* Lanham, MD: Rowman and Littlefield, 2001. xii, 135 p. Cloth $30.00.

SINGER, MING: *Unbounded Consciousness: Qualia, Mind, and Self.* New York: Free Association, 2001. x, 254 p. Cloth $55.00.

SINGER, S. JONATHAN: *The Splendid Feast of Reason.* Berkeley: California UP, 2001. xxiii, 242 p. Cloth $24.95.

SKILLEÅS, OLE MARTIN: *Philosophy and Literature: An Introduction.* New York: Columbia, 2001. viii, 168 p. Paper $19.00.

SLORS, MARC: *The Diachronic Mind: An Essay on Personal Identity, Psychological Continuity, and the Mind-Body Problem.* Norwell, MA: Kluwer, 2001. vii, 234 p. Cloth $82.00.

SMITH, NICHOLAS H.: *Charles Taylor: Meaning, Morals, and Modernity.* Malden, MA: Blackwell, 2002. x, 285 p. Cloth $59.95, paper $24.95.

SOAMES, SCOTT: *Beyond Rigidity: The Unfinished Semantic Agenda of* Naming and Necessity. New York: Oxford, 2002. ix, 379 p. Cloth $35.00.

SOLOMON, MIRIAM: *Social Empiricism.* Cambridge: MIT, 2001. xi, 175 p. Cloth $32.00.

SORENSEN, ROY: *Vagueness and Contradiction.* New York: Oxford, 2001. viii, 200 p. Cloth $35.00.

SOWELL, THOMAS: *A Conflict of Visions: Ideological Origins of Political Struggles.* New York: Basic, 2002. xii, 292 p. Paper $25.95.

STERN, ROBERT: *Hegel and the* Phenomenology of Spirit. New York: Routledge, 2002. xvii, 234 p. Cloth $55.00, paper $14.95.

SULLIVAN, NIKKI: *Tattooed Bodies: Subjectivity, Textuality, Ethics, and Pleasure.* Westport, CT: Praeger, 2001. ix, 203 p. Cloth $62.00.

SULLIVAN, SHANNON: *Living across and through Skins: Transactional Bodies, Pragmatism, and Feminism.* Bloomington: Indiana UP, 2001. ix, 204 p. Cloth $39.95, paper $19.95.

SWINBURNE, RICHARD: *Epistemic Justification.* New York: Oxford, 2001. vi, 262 p. Paper $19.95.

TAYLOR, MARC C.: *The Moment of Complexity: Emerging Network Culture.* Chicago: University Press, 2001. xi, 340 p. Cloth $32.00.

THAU, MICHAEL: *Consciousness and Cognition.* New York: Oxford, 2002. xiii, 280 p. Cloth $39.95.

TRIGG, ROGER: *Philosophy Matters: An Introduction to Philosophy.* Malden, MA: Blackwell, 2002. vii, 157 p. Cloth $54.95, paper $19.95.

UTTAL, WILLIAM R.: *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain.* Cambridge: MIT, 2001. xviii, 255 p. Cloth $39.95.

VAN INWAGEN, PETER: *Ontology, Identity and Modality: Essays in Metaphysics.* New York: Cambridge, 2001. ix, 261 p. Cloth $64.95, paper $22.95.

VAN MILL, DAVID: *Liberty, Rationality, and Agency in Hobbes's* Leviathan. Albany: SUNY, 2001. xii, 253 p. Cloth $59.50, paper $19.95.

VILLA, DANA: *Socratic Citizenship.* Princeton: University Press, 2001. xvi, 370 p. Cloth $65.00, paper $17.95.

VOELLER, CAROL: *The Metaphysics of the Moral Law: Kant's Deduction of Freedom.* New York: Garland, 2001. xvii, 175 p. Cloth $60.00.

VOKEY, DANIEL: *Moral Discourse in a Pluralistic World.* Notre Dame, IN: University Press, 2001. x, 373 p. Cloth $45.00.

VON DIETZE, ERICH: *Paradigms Explained: Rethinking Thomas Kuhn's Philosophy of Science.* Westport, CT: Praeger, 2001. x, 182 p. Cloth $64.00.

WALTERS, GREGORY J.: *Human Rights in an Information Age: A Philosophical Analysis.* Toronto: University Press, 2001. xx, 335 p. Cloth $65.00.

WARREN, DANIEL: *Reality and Impenetrability in Kant's Philosophy of Nature.* New York: Routledge, 2001. xiv, 99 p. Cloth $60.00.

WATKINS, JOHN: *Human Freedom after Darwin.* Chicago: Open Court, 1999. xi, 348 p. Cloth $49.95, paper $24.95.

THE JOURNAL OF
PHILOSOPHY

The Journal is pleased to announce to our readers and subscribers the launch of its new website, which will have the pages listed below. The address is: *www.journalofphilosophy.org* or simply *www.jphil.org*.

~•~

PURPOSE & HISTORY
CONTACTS AND ADDRESSES
BOARD OF TRUSTEES
EDITORIAL BOARD
MEMOS TO AUTHORS
SUBSCRIPTIONS & ADVERTISING

RECENT & FORTHCOMING ISSUES
ELECTRONIC ISSUES
INDEX 1904–PRESENT
RIGHTS AND PERMISSIONS
LECTURES & CONFERENCES
PHILOSOPHY LINKS

~•~

---

from DE GRUYTER...

Current Issues in Theoretical Philosophy
Volume 1: What is Truth?
Richard Schantz, editor

2001. vi + 340 pages. Cloth $71.15 ISBN 3-11-016441-8

In this collection of original papers, leading international authorities turn their attention to one of the most important questions in theoretical philosophy: What is Truth? To arrive at an answer, two further questions need to be addressed in the following context: 1) Does truth possess any essence, any inner nature? and 2) If so, what does this nature consist of?

The present discussion focuses on the antagonism between substantial or sound theories of truth, with the correspondence theory taking the lead, and deflationist or minimalist views, which have been commanding an increasing amount of attention in recent years. Whereas substantial theories proceed from the premise that truth has an essence, and that therefore the objective is to discover this essence, the challenge presented by deflationism is to abandon this very premise.

Authors: W. P. Alston, D. M. Armstrong, R. Boyd, M. Devitt, R. Schantz, R. Brandom, D. Grover, P. Horwich, M. Williams, M. David, W Künne, C. McGinn, P. van Inwagen, A. Gupta, J. Hintikka, M. Dummett, L. B. Puntel, J. Rosenberg, R. Walker, D. Wiggins

Richard Schantz is Professor of Philosophy at Siegen University, Germany.

Price is subject to change

Walter de Gruyter, Inc.
200 Saw Mill River Road, Hawthorne, NY 10532 • Tel. (914) 747-0110 • Fax (914) 747-1326
E-mail: cs@degruyterny.com • Visit Our Website: www.degruyter.com

PHILOSOPHY OF SCIENCE

The official Journal
of the Philosophy
of Science
Association

Editor-in-Chief
Noretta Koertge
Indiana University

Associate Editors
W. Michael Dickson
Michael Friedman
Elisabeth A. Lloyd

Published quarterly, plus supplemental volume of the Proceedings of the Biennial Meeting of the Philosophy of Science Association.

Purpose and scope

Founded in 1934, ***Philosophy of Science*** has become the authoritative publication concerned with philosophical inquiry into the theory and practice of natural and social science. Adhering to the strictest standards in its selection of essays, discussion articles and book reviews, the ***Journal*** brings together the best scholarship on the methodological, ontological, epistemological, anthropological, and ethical premises and implications of the various scientific disciplines. In addition to new developments in traditional areas, ***Philosophy of Science*** presents ground-breaking work on conceptual issues that arise in new areas of scientific inquiry, such as bioinformatics, cognitive ethology, and neuroscience, as well as discussions of policy issues connected with the changing role of science in our society.

Intended Audience

Philosophy of Science is not just for philosophers! It will be of interest to natural scientists searching for new perspectives on their fields, epistemologists, and cognitive scientists who study the organization of knowledge, and social scientists and humanists who approach science as a cultural phenomenon.

Membership in the Philosophy of Science Association includes the journal ***Philosophy of Science*** *at no additional charge. Individual subscriptions to the* journal *and membership in the PSA are concurrent*

Year 2001 PSA membership dues/subscription rates: Institutions-A (Australia, Canada, Hong Kong, Japan, New Zealand, Western Europe, and U S.) $102.00; Institutions-B (all other countries) $51 00. Members (income less than $40,00) $57.00, Members (income $40,000-$49,999) $67.00; Members (income $50,000-$59,999) $77.00; Members (income more than $60,000) $87 00; Members-B (all other countries) $25.00; Student members (with copy of valid ID) $25 00. Canadians, please add 7% GST plus postage. Institutions outside USA, please add $6 postage. MasterCard and Visa orders accepted (please include account number, expiration date, signature, and phone number); credit card orders may be sent via fax to (773) 753-0811or (877) 705-1879. Send your order with check, credit card information, or purchase order to

The University of Chicago Press
Journals Division, P.O. Box 37005, Chicago, IL 60637
Order online at www journals.uchicago.edu/PHILSCI

2/02

New Philosophy Titles from

HB Humanity Books

KARL JASPERS ON PHILOSOPHY OF HISTORY AND HISTORY OF PHILOSOPHY

Edited by Joseph W. Koterski, S.J., and Raymond J. Langley

This volume contains the first English translations of texts by Karl Jaspers on Aquinas, Kant, Hegel, and Kierkegaard. In addition, interpretive and critical essays by renowned scholars assess Jaspers' contributions to the history of philosophy and the philosophy of history. This stimulating collection will complement college and university libraries and be of value to anyone interested in comparative Eastern and Western philosophy, the history of ideas, or the interrelations of biography and history.

300 pp • ISBN 1-59102-002-6 • HC $65

PASSION AND VIRTUE IN DESCARTES

Edited by Byron Williston and André Gombay

In the period following publication of the *Meditations*, Descartes was extremely active in attempting to develop a comprehensive ethics, rooted in his analysis of human passions. The present volume is the first collection of original essays—all by leading Descartes scholars—dealing specifically with this aspect of Descartes' thinking as presented in the *Passions of the Soul*. Since in the *Passions* Descartes provides us with the most sophisticated discussion of the composite human being anywhere in his writings, this volume is an indispensable aid to understanding the full scope of Cartesian anthropology.

285 pp • ISBN 1-59102-005-0 • HC $59

DEFINING DIALOGUE

From Socrates to the Internet

Geoffrey Rockwell

This original cross-disciplinary work examines the crucial role of dialogue in philosophy from the oral dialogues of Socrates, through the written dialogues of Plato, Cicero, Lucian, Valla, Hume, and Heidegger, to the present ubiquitous form of dialogue on the Internet. Geoffrey Rockwell's main point is that in dialogue, be it oral, written, or electronic, there is a common mode of persuasion at work. The dialogue is an orchestrated event meant to be overheard. It is a genre suited to presenting how people discuss ideas, how positions are related to character, and surveying positions that can be taken on a subject. In a world increasingly connected by the Internet, there is no more appropriate genre for study.

225 pp • ISBN 1-57392-954-9 • HC $49

For exam copies contact editorial@prometheusbooks.com
Call toll free (800) 421-0351 • fax (716) 691-0137
www.prometheusbooks.com

New this Spring from Indiana...

BECOMING JOHN DEWEY
Dilemmas of a Philosopher and Naturalist
Thomas C. Dalton
A new intellectual biography of John Dewey that emphasizes his ties with psychology and the sciences
cloth $45.00

SOCIAL EPISTEMOLOGY
Second Edition
Steve Fuller
Foreword by Thomas Nickles
A new edition of Fuller's acclaimed book about the way knowledge is organized.
cloth $49.95 paper $21.95

A SHORT HISTORY OF AFRICAN PHILOSOPHY
Barry Hallen
A concise and comprehensive history of African philosophy
cloth $29.95 paper $14.95

SKEPTICAL ENVIRONMENTALISM
The Limits of Philosophy and Science
Robert Kirkman
A skeptical and practical look at environmental philosophy and its limitations.
cloth $39.95 paper $19.95

MAIN CURRENTS IN WESTERN ENVIRONMENTAL THOUGHT
Peter Hay
A comprehensive, readable overview of three decades of western debates on environmental issues.
cloth $59.95 paper $29.95

THE VOICE OF HARRIET TAYLOR MILL
Jo Ellen Jacobs
Innovative biography of the woman who wrote *On Liberty.*
cloth $35.00

THE LIVES OF THINGS
Charles E. Scott
An animated reflection on the capacity of things to inspire and astonish.
Studies in Continental Thought
cloth $49.95 paper $19.95

DERRIDA AND HUSSERL
The Basic Problem of Phenomenology
Leonard Lawlor
The first systematic study of Derrida's writings on Husserl.
Studies in Continental Thought
cloth $49.95 paper $19.95

NATIVE PRAGMATISM
Rethinking the Roots of American Philosophy
Scott L. Pratt
Explores the connections between American pragmatism and Native American thought
cloth $44.95 paper $19.95

at bookstores
INDIANA
University Press

Orders: 1-800-842-6796 • iupress.indiana.edu

Best in Philosophy from Cambridge

New Series...

Contemporary Philosophy in Focus

Daniel Dennett

Andrew Brook and
Don Ross, Editors

0-521-80394-2	Hardback	$60.00
0-521-00864-6	Paperback	$20.00

Robert Nozick

David Schmidtz, Editor

0-521-78226-0	Hardback	$60.00
0-521-00671-6	Paperback	$20.00

Now in paperback...

From Chance to Choice

Genetics and Justice

Allen Buchanan, Dan Brock,
Norman Daniels, and Daniel Wikler

0-521-66977-4	Paperback	$23.00

The Cognitive Basis of Science

Peter Carruthers, Stephen Stich, and
Michael Siegal, Editors

0-521-81229-1	Hardback	$70.00
0-521-01177-9	Paperback	$25.00

The Cambridge Companion to Newton

I. Bernard Cohen and
George E. Smith, Editors

Cambridge Companions to Philosophy

0-521-65177-8	Hardback	$55.00
0-521-65696-6	Paperback	$20.00

The Cambridge Companion to Gadamer

Robert J. Dostal, Editor

Cambridge Companions to Philosophy

0-521-80193-1	Hardback	$65.00
0-521-00041-6	Paperback	$23.00

Practical Reasoning in a Social World

How We Act Together

Keith Graham

0-521-80378-0	Hardback	$55.00

The Myth of Morality

Richard Joyce

Cambridge Studies in Philosophy

0-521-80806-5	Hardback	$60.00

Nietzsche, Biology and Metaphor

Gregory Moore

0-521-81230-5	Hardback	$55.00

Nietzsche: *Beyond Good and Evil*

Friedrich Nietzsche

Cambridge Texts in the History of Philosophy

0-521-77078-5	Hardback	$38.00
0-521-77913-8	Paperback	$14.00

Disseminating Darwinism

The Role of Place, Race, Religion, and Gender

Ronald L. Numbers and
John Stenhouse, Editors

0-521-01105-1	Paperback	$22.00

Wittgenstein's Tractatus

A Dialectical Interpretation

Matthew B. Ostrow

0-521-80936-3	Hardback	$55.00
0-521-00649-X	Paperback	$20.00

Real Ethics

Reconsidering the Foundations of Morality

John M. Rist

0-521-80921-5	Hardback	$65.00
0-521-00608-2	Paperback	$23.00

The Nature of Consciousness

Mark Rowlands

0-521-80858-8	Hardback	$60.00

The Theory of Moral Sentiments

Adam Smith
Knud Haakonssen, Editor

Cambridge Texts in the History of Philosophy

0-521-59150-3	Hardback	$65.00
0-521-59847-8	Paperback	$23.00

Heidegger's Later Philosophy

Julian Young

0-521-80922-3	Hardback	$55.00
0-521-00609-0	Paperback	$20.00

Available in bookstores or from

CAMBRIDGE
UNIVERSITY PRESS

800-872-7423
us.cambridge.org/philosophy

THE JOURNAL OF

PHILOSOPHY

SUBSCRIPTIONS

2002

Students $20.00
Individuals $35.00
Institutions $75.00
Postage outside the U.S. $15.00

SPECIAL OFFER

one complimentary copy of any of the following
special issues with each new individual subscription

2000	*Causation*	David Lewis, Ned Hall, John Collins, L.A. Paul, Jonathan Schaffer
1998	*Having the World in View: Sellars, Kant, and Intentionality*	John McDowell
1995	*Reconciliation through the Public Use of Reason*	Jürgen Habermas
	Reply to Habermas	John Rawls
1994	*Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind*	Hilary Putnam
1990	*The Structure and Content of Truth*	Donald Davidson
1985	*Well-Being, Agency, and Freedom*	Amartya Sen
1980	*Kantian Constructivism in Moral Theory*	John Rawls
1977	*Teleology Revisited*	Ernest Nagel
1968	*Ontological Relativity*	W.V. Quine

All issues above are also available for $8.00 plus $3.00 for shipping/handling ($5.00 for international and Canadian orders). Postage for each added issue, $1.00. Send check or money order to:

Journal of Philosophy
709 Philosophy Hall
1150 Amsterdam Avenue, MC 4972
Columbia University
New York, NY 10027
(212) 666-4419
www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

9/8/02

VOLUME XCIX, NUMBER 5
MAY 2002

page
223 *Moral Worth* Nomy Arpaly
246 *What Is the Problem of Coherence and Truth?* Erik J. Olsson

273 NEW BOOKS: ANTHOLOGIES

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2002

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL'S terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2002 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME XCIX, NO. 5, MAY 2002

9187

MORAL WORTH*

Sometimes a person does the right thing without earning our admiration. The reader of Gustave Flaubert's *Madame Bovary*[1] is told of the protagonist:

> She sewed clothes for the poor, she sent wood to women in childbirth; and on coming home one day, Charles found three tramps eating soup in the kitchen. Her little girl, whom her husband had sent back to the nurse during her illness, returned home. She wanted to teach her to read; even Berthe's crying no longer irritated her. She was resigned, universally tolerant. Her speech was full of elevated expressions. She would say: "Is your stomach-ache any better, my angel" (*ibid.*, p. 155)?

Flaubert does not admire Madame Bovary for her charity, referring to her as "indulging" in "excessive charity." But why? It is hardly morally wrong for her to do so much. But Madame Bovary's good works are taken up, briefly, after a long illness brought about by the traumatic end of her first extramarital affair, and she loses her interest in orphans as soon as her next lover comes along.

Madame Bovary's actions appear motivated by a mere infatuation with morality. Perhaps her actions are "indulgent" in that, even though she desires to be moral and performs her good works because they are moral, she has an ulterior motive of sorts. Like the romantic who is in love not with her lover but with love, Madame Bovary seems to be in love not so much with morality as with the romance of morality. To desire to become the kind of person who cares about

---

* I would like to thank my friends at the University of Michigan for tremendous help in the development of this article. I am also grateful to Michael Bratman, Rachel Cohon, Russell Hardin, Barbara Herman, Kim Kane-Maguire, Timothy Schroeder, and George Sher for their comments on various stages of this work.

[1] Paul de Man, trans. (New York: Norton, 1965).

0022-362X/02/9905/223–45

morality may be a worthy motive, but something about Madame Bovary's dwelling upon the long-trimmed trains of these ladies' long gowns, something about the quasi-sexual passion with which she kisses the image of Christ seems suspicious. Why the same actions prompt us to praise—or blame—some agents much less than others is what I shall call the *question of moral worth*.

I. A FEW CLARIFICATIONS

The moral worth of an action is the extent to which the agent deserves moral praise or blame for performing the action, the extent to which the action speaks well of the agent. I shall speak interchangeably of a *morally praiseworthy action* and an *action which has positive moral worth*. In deviation from the Kantian use of the term 'moral worth', I shall also speak interchangeably of an action with *negative moral worth* as an action for which the agent is blameworthy and as a *morally blameworthy action*.

Obviously, the extent to which an agent deserves praise or blame for her action depends substantially on the action's *moral desirability*: whether it is right or wrong, or how grave a wrong it is, or whether it is the best possible action. But two actions that are equal in moral desirability may be of different moral worth. To give a simple example, two people may donate equally to Oxfam, but one of them may do so out of compassion, while the other does so purely at the urgings of her accountant. Even if the two agents' deeds are equally morally desirable, it is not true that both deserve the same praise. Similarly, a person rude to her colleagues due to the stress caused by grave news may merit less blame than a genius rude to her colleagues due to her indifference to "lesser minds."

While this distinction between *moral desirability* and *moral worth* appears trivial, it is often ignored. One hears that "Kantians are concerned with the motives for our actions, while utilitarians only care about the results," but paying attention to the distinction between moral desirability and moral worth reveals the falsity of the classroom cliché. When it comes to moral desirability, Kantians are interested in whether the action, under a certain description, is permitted by a universal law. Recall Immanuel Kant's prudent grocer, who prices his merchandise fairly because a reputation for honesty tends to increase his profit. Despite the prudent grocer's unimpressive motive, Kant never denies that the grocer does *the right thing*. In this sense, Kantians care about more than motives alone. On the other hand, when it comes to moral worth, utilitarians are free to be as concerned with motives as Kantians. In fact, just as John Stuart

Mill[2] repeats the claim that the rightness or wrongfulness of actions has nothing to do with the agent's motives, he also repeats the claim that motives are relevant to our moral evaluation of the agent:

> ...the motive has nothing to do with the morality of the action, though much with the worth of the agent.... The motive, that is, the feeling which makes him will so to do, if it makes no difference in the act, makes none in the morality [of the act]; though *it makes a great difference in our moral estimation of the agent* (*ibid.*, pp. 17-18; italics added).

As the moral worth of an action is the "estimation" merited by the agent for the action, it is not implausible to think of Mill as allowing motives to be relevant to the moral worth of actions, even if they have nothing to do with rightness or wrongfulness. More important, putting interpretive issues aside, it is quite consistent to view the moral desirability of actions as depending entirely on their consequences, and the moral worth of actions as depending partly on motive. One can believe that giving to charity is desirable because it promotes happiness, but that an agent giving to charity out of a desire for this end merits more praise than does her counterpart who is merely concerned with her tax situation.[3]

II. RESPONSIVENESS TO MORAL REASONS

Consider again the prudent grocer. One need not know the details of Kant's discussion to agree that a grocer motivated solely by concern for profit is not particularly praiseworthy for pricing fairly. One is happy the grocer does the right thing, but feels it is a mere accident, for which one is not inclined to give the grocer moral credit. But what, exactly, makes it accidental? It is *not* simply the fact that the profit motive does not reliably produce moral actions. We can imagine a world in which some invisible hand makes the profit motive reliably produce morally right actions, place Kant's grocer in that world, and still not free ourselves from the sense that there is something accidental in his acting well. It is accidental in the same way it is accidental that a person reading Vladimir Nabokov's *Lolita* for the love of scandal reads an aesthetically superior book. The scandal lover is attracted to *Lolita* for reasons that are of no interest to the aesthetician, and Kant's grocer is attracted to fair pricing for reasons that are of no interest to the ethicist. The salient feature of Kant's case is that the grocer's action does not stem from any responsiveness on his part to moral reasons. In pricing fairly, the grocer acts for a reason that has nothing to do with morality, or with the features of his action which makes it morally right. His *reasons for action* do not correspond to the action's *right-making features*, suggesting:

---

[2] *Utilitarianism* (Indianapolis: Hackett, 1979).

[3] Note that the moral worth of an action is the extent to which the agent deserves praise or blame for the action, not the extent to which the agent should be morally praised or blamed for it. The purpose of this work is to capture the conditions under which praise or blame is *warranted*, not those under which it is *required*. See my "Hamlet and the Utilitarians," *Philosophical Studies*, XCIX (2000): 45-57.

> *Praiseworthiness as responsiveness to moral reasons*: for an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons, that is, the reasons making it right.

Kant appeals to this idea when he argues that only the good will is necessarily good, and Aristotle appeals to it when he explains that defending one's city would not be virtuous if motivated by desire for fame. Thus, it may look tempting, or even trivial, to move to the more elegant Kantian claim that morally praiseworthy actions are all and only those right actions performed "from duty," or the venerable Aristotelian view that to act virtuously, one needs to perform fine actions "for the sake of the fine." But as these claims are often understood, this would be an error. Usually, "acting from duty" and "acting for the sake of the fine" are taken to indicate acting for reasons *believed* or *known* to be moral reasons; but it can be shown that moral worth is fundamentally about acting for moral reasons, not about acting for reasons believed or known to be such.

III. MORAL RESPONSIVENESS VERSUS CONCERN FOR MORALITY

Consider the Kantian claim that all and only morally praiseworthy right actions—or, in his language, all and only dutiful actions that "have moral worth"—are performed from duty. In her "On the Value of Acting from the Motive of Duty," Barbara Herman[4] phrases the doctrine of the motive of duty in the following way:

> For a motive to be a moral motive, it must provide the agent with an interest in the general rightness of his actions. And when we say that an action has moral worth, we mean to indicate (at the very least) that the agent acted dutifully from an interest in the rightness of his action: an interest that therefore makes its being a right action the non-accidental effect of the agent's concern (*ibid.*, p. 6).

But for a right action to have (positive) moral worth, it is neither sufficient nor necessary that it stem from the agent's interest in the rightness of his action. To see that it is not *sufficient*, one can look at cases in which an agent does the right thing out of concern for doing the right thing, and still the fact that he did the right thing appears accidental. This can happen when the agent has a mistaken view of morality. Consider:

---

[4] In *The Practice of Moral Judgment* (Cambridge: Harvard, 1993), pp. 1-22.

> "The Extremist": after the assassination of Yitzhak Rabin, some Jewish extremists expressed the opinion that the murder was horrible mostly because it involved a Jew killing a Jew. Imagine for a moment that Ron is such an extremist, believing deeply that killing a person is not generally immoral, but that killing a fellow Jew is. Ron would very much like to kill Tamara, but he refrains from doing so, because he wants to do the right thing, and he believes the right thing to do is not to kill Jews like Tamara.

Here Ron does the right thing—refrain from killing Tamara—because he wants to do the right thing and believes, rightly, that refraining from killing would be the right thing to do. He does not, however, do the right thing for the relevant moral reasons, Tamara's ethnicity being morally irrelevant—hence the impression that it is merely *accidental* that Ron does the right thing. Ron is not morally praiseworthy for his inaction, for he is unmoved by the morally relevant reasons, but rather by reasons he mistakenly believes to be morally relevant.

The story of Ron shows that for an agent to be morally praiseworthy for acting well it is not sufficient that she act on a desire to do what is right. One might suggest instead that one is morally praiseworthy for acting well if and only if one acts on a desire to do what is right and has moral knowledge, or knowledge of the virtues. Perhaps this formulation would better capture Kant's view, and it certainly appears to be Aristotle's. But even if acting on a desire to do what is right assisted by knowledge of the right is sufficient for granting one's good deeds moral worth, it is not necessary for it. Consider cases in which a person does the right thing for moral reasons, but does not in any way act out of a desire to do the right thing. It is exactly ignorance of the virtues, or lack of moral knowledge, that can lead a person to this state. The same ignorance that led the extremist to mistake racist reasons for moral reasons can sometimes cause a person to mistake moral reasons—for which he acts—for something else.

Consider first what I have called *inverse akrasia*,[5] a term referring to the phenomenon in which an agent does the right thing, but does so against her best judgment. Donald Davidson[6] points out that, contra Aristotle, akrasia does not always involve an agent's doing something she should not do—it need only involve an agent's doing something that she thinks she should not do. If we were to believe that the average person *should* brush his teeth every night, Davidson's well-known case of the man who decides not to brush his teeth but brushes them anyway would be a case of inverse akrasia.[7] Some cases of inverse akrasia are quite clearly cases in which an agent accidentally does something that happens to be right, that is, cases of right action without moral worth. But some cases of inverse akrasia are also cases of morally praiseworthy action.

---

[5] "Praise, Blame, and the Whole Self," *Philosophical Studies*, XCIII (1998): 161-88.

[6] *Essays on Actions and Events* (New York: Oxford, 1980), see p. 31.

In Mark Twain's *The Adventures of Huckleberry Finn*,[8] Huckleberry befriends Jim and helps him escape from slavery. While Huckleberry and Jim are together on their raft, Huckleberry is plagued by what he calls "conscience." He believes, as everyone in his society "knows," that helping a slave escape amounts to stealing, and stealing is wrong. He also believes that one should be helpful and loyal to one's friends, but loyalty to friends is outweighed by some things, such as property rights, and does Miss Watson, Jim's owner, not have property rights? Hoping to find some excuse not to turn Jim in, Huckleberry deliberates. He is not very good at abstract deliberation, and it never occurs to him to doubt what his society considers common sense. Thus, he fails to find a loophole. "What has poor Miss Watson done to me," he berates himself, "that I can see her nigger go away and say nothing at all?" Having thus deliberated, Huckleberry resolves to turn Jim in, because it is "the right thing." But along comes a perfect opportunity, and he finds himself psychologically unable to do it. He accuses himself of being a weak-willed boy who has not "the spunk of a rabbit," and decides that being moral is such a hard and thankless task that he might as well give up and remain a bad boy. Obviously, Huckleberry does the right thing, but does he merit praise? The answer to this question depends on our reconstruction of his motives, but, on at least some, Huckleberry is morally praiseworthy for his action. If Huckleberry were to help Jim because of the operation, within himself, of some purely atavistic mechanism—akin, perhaps, to the human tendency to favor animals with infantile appearance

---

[7] Other cases of inverse akrasia are described in Harry Frankfurt, *The Importance of What We Care About* (New York: Cambridge, 1988); Alasdair MacIntyre, "Is Akratic Action Always Irrational?," in Owen Flanagan and Amelie Oksenberg Rorty, eds., *Identity, Character and Morality* (Cambridge: MIT, 1990), pp. 379-400; and Robert Audi, "Weakness of Will and Rational Action," *Australasian Journal of Philosophy*, LXVIII (1990): 271-81. The case of Neoptolemus as described in the *Nicomachean Ethics* also counts as a case of inverse akrasia, so long as we have a Davidsonian and not an Aristotelian concept of akrasia in mind.

[8] *The Works of Mark Twain*, Volume 8 (Berkeley: California, 1988).

(big eyes, and the like)—we would not regard his action as morally praiseworthy, because he would not be acting for reasons at all. This is, perhaps, what Kant thinks about when he talks of acting out of "mere inclination," and it is the interpretation favored by Jonathan Bennett,[9] who sees Huckleberry as merely squeamish, soft-hearted, unable to see a man in chains. In this interpretation, Huckleberry is a racist boy who accidentally does something good. There is a version of the story, however, in which Huckleberry is morally praiseworthy for his action.[10] In this scenario, his visceral experience of black people is inconsistent with his "official" racist views. There are people who sport liberal views but cross the road when a person of a different race appears. Huckleberry, on this reading, is the opposite: racist in conscious opinion but viscerally more egalitarian. This discrepancy widens during the time he spends with Jim. Talking to Jim and interacting with him, Huckleberry constantly perceives data (never deliberated upon) that amount to the impression that Jim is a full person, just like Huckleberry himself. While he never deliberates on his perceptions,[11] they prompt him increasingly to act toward Jim as a friend. Twain makes it very easy for the boy to perceive that Jim is very similar to him: Jim shares Huckleberry's language, knowledge, ignorance, and superstitions; and all-in-all it does not take the genius of Mill to see that there is no particular reason to think of one of them as inferior to the other. That Huckleberry begins to perceive Jim as a fellow human being is suggested when Huckleberry finds himself, to his surprise, *apologizing* to Jim—an action unthinkable in a society treating black men as subhuman. Contra Rosalind Hursthouse,[12] my point is not simply that Huckleberry does not have the belief that his action is moral on his mind while he acts. He does not have the belief that what he does is right *anywhere* in his head—this moral insight is exactly what eludes him. Yet when the opportunity comes to turn Jim in, and Huckleberry experiences a strong reluctance to do so, his reluctance is to a large extent the result of the fact that he has come to perceive Jim as a person, even if his conscious mind has not yet come to reflective awareness of this. To the extent that Huckleberry is reluctant to turn Jim in because of Jim's personhood, he *is* acting for morally significant reasons. This is so even though Huckleberry knows neither that these are the right reasons nor that that he is acting from them. On this reading, he is not a bad boy who has accidentally done something good, but a good boy with imperfect knowledge.

---

[9] "The Conscience of Huckleberry Finn," *Philosophy*, xlix (1974): 123-34.

[10] For an alternative, sophisticated treatment of Huckleberry Finn, see Thomas Hill, "Four Conceptions of Conscience," *Nomos*, lx (1998): 13-52. Julia Driver also takes for granted that Huckleberry's action is meant to be understood as praiseworthy, but her account of this praiseworthiness is completely independent of Huckleberry's motives and reasons; see her "The Virtues and Human Nature," in Roger Crisp, ed., *How Should One Live? Essays on the Virtues* (New York: Oxford, 1996), pp. 111-30.

[11] There is nothing very unusual about perceiving fairly sophisticated truths without perceiving that one is perceiving them—as when the confession of a cheating spouse is surprisingly unsurprising. I argue for this conclusion in detail in "On Acting Rationally against One's Best Judgment," *Ethics*, cx (2000): 488-513, and it is often presupposed by cognitive scientists.

[12] *On Virtue Ethics* (New York: Oxford, 2000).

Huckleberry is not an isolated case. Inverse akrasia is not unlike a more common sort of behavioral inconsistency: the person whose explicit moral or political views are terribly wrong, but who in practice "cannot hurt a fly." We all have friends, family members, or acquaintances of this sort: the likes of a student who, waving his copy of Ayn Rand's *Atlas Shrugged* in one's face, preaches that one should be selfish, and then proceeds to lose sleep generously helping his peers. Were philosophers right in believing that only actions following from one's conscious moral principles are done for moral reasons, we should view these people as bad people with moral luck. More commonly, however, we treat these people as fundamentally good people who happen to be incompetent abstract thinkers. While such people may be baffling—note the fascination with the character of Oskar Schindler—they are as commonplace as their opposite numbers, people with wonderful convictions who act immorally. The idea that we can sometimes act for moral reasons without knowing that we act for moral reasons is not strange when posed against the background of epistemology and psychology, where many have maintained that we can know without knowing that we know, believe without believing that we believe, or act for a reason without knowing that we act for a reason. It is only strange when posed against certain traditions in ethics and action theory, wherein we are used to viewing people as divided neatly into the faculty of reason and that shady realm of emotion, inclination, and instinct. I argue against this picture elsewhere;[13] here, I would just like to point out the following: on the reason/inclination picture, an agent who is pulled into action by "inclination" can be blamed only indirectly—for failing to restrain his desires and emotions properly—and can be praised only indirectly—for managing to train his desires and emotions so that they appear in the right time and place. One may attempt, with difficulty, to force the plethora of different cases in which one condemns the viscerally racist liberal or the child-beating Christian into a picture in which it is only a lack of self-control that one condemns. But Huckleberry Finn, Oskar Schindler, and our *Atlas Shrugged*-toting friend are obviously not praiseworthy for any kind of self-training or character-building on their parts. They are praiseworthy because some of their moral common sense—their responsiveness to moral reasons—is intact.

---

[13] "On Acting Rationally against One's Best Judgment."

IV. BLAME AND MORAL UNRESPONSIVENESS

To recapitulate: some people do the right thing by accident, while others do the right thing in response to moral reasons. Those who do the right thing in response to moral reasons are those who are morally praiseworthy for their actions. It stands to reason, then, that something similar would be true of people who do the wrong thing. Sometimes, the fact that one did the wrong thing appears to be accidental, while at other times it seems to stem from what can be called "ill will," or from a deficiency of good will. If good will is responsiveness to moral reasons, deficiency in good will is insufficient responsiveness to moral reasons, and ill will is responsiveness to sinister reasons—reasons that, in their essence, conflict with morality. Imagine, for example, that Jeanne is very rude to Joseph and hurts his feelings. Many scenarios can be imagined which fill in her motives. Perhaps she comes from an aggressive culture in which "shut up" is a commonly used phrase, and has not met people who find it offensive. In this case, the fact that she did something hurtful is, in a clear sense, accidental. Perhaps, on the other hand, she has no such excuse, and she acts the way she does because she desires to vent the tensions of a long day by saying exactly what comes to her mind, which happens to be offensive to Joseph in an obvious way. In this case, she is blameworthy, because her action indicates a failure to respond to morally relevant considerations—she should be motivated by the fact that Joseph is likely to be hurt by what she says, but she is unmoved. A third possibility is that she is rude to Joseph because she enjoys inflicting suffering on others and wishes to hurt Joseph. In this scenario, she is even more blameworthy than in the previous one, as her action not only expresses a deficiency of good will, but also expresses ill will. There is nothing about the desire to vent one's feelings which essentially conflicts with morality. Like the more celebrated motives of love and money, it can sometimes conflict with morality, while at other times it leads to morally good or neutral actions. On the other hand, a desire to inflict suffering for its own sake is essentially in conflict with morality. To do something purely because it would inflict suffering on a fellow human being is to act for sinister or "antimoral" reasons. Other things being equal, a person is more blameworthy for a given wrong if she acts out of ill will (from sinister reasons) than she would be if she were to act out of a lack of good will (out of neutral reasons, while ignoring moral reasons to the contrary). This does not, however, imply that a sadist or a proponent of a sinister ideology is always a worse person than an opportunist. A chilling, profound indifference to moral reasons, the kind compatible with, say, killing for profit, is much worse than a predilection for the mildly immoral, such as a taste for unnerving one's underlings.

Consider again the person who fails to respond to relevant moral factors. There seems to be more than one way one can lack this responsiveness. Sometimes, one is deficient in moral perception—blind, as it were, to some moral factors. Some people, for example, do not seem to be able to grasp the idea of personal autonomy. Imagine here the sort of parent who not only fails to see that anything could be wrong with her extremely paternalistic treatment of her adult son, but cannot understand the concept of "paternalism" even after it has been explained over and over. ("But if no one tells you what to do, how will you know?" she asks in bewilderment.) More often, however, one perceives all the morally relevant features of a situation, but is not sufficiently moved by them. One need not be an amoralist for this to happen; it is enough that one be a more common, though much less discussed figure—the human, all-too-human person who cares about morality, but *not very much*. Mary may realize that, if she does not promptly return a friend's book, she will break a promise and cause unnecessary distress; and she may still not send it, because it is freezing outside or because it would be nice if she finished the novel she is writing before her birthday, and so she cannot really spare the time to make a special trip. Not being an amoralist, it is not the case that she is indifferent to the institution of promising or other relevant moral factors. She cares about these things—enough to feel morally guilty for her course of action and enough not to commit more serious breaches of trust, but not enough to motivate her to go to the post office. The first kind of morally unresponsive person is analogous to the person who does not notice it when his clothes do not match. The second kind is analogous to the person who notices very well when his clothes are unmatched, but does not care about appearances enough to make sure that he is never stuck with nothing to wear but his orange shirt and purple pants. Deficiency of perception and deficiency of motivation are, admittedly, hard to tell apart. Many people have wondered if a spouse does not *see* the dust on the floor or does not *care* if it is there, and the diagnosis is hard because there is often a combination. Similar things apply to the person of mediocre moral sensibilities. A person who does not care much about morality may not give much thought to some things to which a more morally concerned person would pay more attention, and thus in fact may be less competent in perceiving other people's feelings, putting herself in their shoes, and so on.[14]

IV DEGREES OF MORAL CONCERN

The above discussion of moral unresponsiveness has brought to the foreground the idea that, given that a person is motivated by moral (or antimoral) reasons, she can be motivated by them more or less powerfully. Let us go back, for a moment, to the realm of praise. Kant's prudent grocer and his ilk demonstrate to us that moral praise is warranted for the agent who acts for moral reasons, as opposed to others. But suppose we know that an agent acts for moral reasons. What bearing should the amount of concern that he has for moral considerations have on our assessment of his action's moral worth? I suggest:

> *Praiseworthiness as responsiveness to moral reasons (revised version)*: for an agent to be morally praiseworthy for doing the right thing is for her to have done the right thing for the relevant moral reasons—that is, in response to the features that make it right (the *right reasons* clause); and an agent is more praiseworthy, other things being equal, the stronger the moral concern that has led to her action (the *concern* clause). Moral concern is to be understood as concern for what is in fact morally relevant and not as concern for what the agent takes to be morality.[15]

I take concern to be a form of desire: for a person to be concerned with morality is for her to have an intrinsic desire that people (herself included) do that which is, in fact, moral. Much work has been devoted to the study of desire, and very different views have been defended by philosophers of mind.[16] Giving a full account of concern in general or moral concern in particular would require choosing a theory of desire, which is not my aim here.[17] What I shall do instead is, first, sharpen our intuitive idea of concern by making a few points about what depth of concern does *not* amount to, and, second, to point out what I take to be important markers of depth of concern.

---

[14] By distinguishing, at least at the phenomenal level, the motivational from the perceptual side of responsiveness, I do not preclude the possibility that failing to care about morality is always irrational (after all, failing to care about one's health or bank account may be irrational as well), nor even the view that moral factors when noticed are always *somewhat* motivating. All that I assume here is that two agents can be motivated by the same moral reasons exactly—but one of them may be motivated by them much more than the other.

[15] One can, if one is so inclined, apply Michael Smith's discussion of the distinction between concern for morality *de re* and *de dicto* in *The Moral Problem* (Cambridge: Blackwell, 1994).

[16] For a sense of the diversity of prevailing views, see, for example, Fred Dretske, *Explaining Behavior* (Cambridge: MIT, 1988); Robert Stalnaker, *Inquiry* (Cambridge: MIT, 1984); Galen Strawson, *Mental Reality* (Cambridge: MIT, 1994)

Strength of concern does not amount to a type of *reflective endorsement*. Erica and I may reflectively endorse the same kind of political action, but she may be more concerned with it than I am, which may explain why she is at a demonstration while I am writing. It is also natural to say that I am less *committed* to political action than Erica is, and this may tempt some readers to think of caring in terms of *commitment*—reflective endorsement with some sort of emotional back-up, perhaps. This would also be misleading, as we may deeply care about things that we do not reflectively endorse at all. Lisa may care deeply about Todd and her relationship with him even though she believes she should not do so, or even though she is utterly unaware of her deep concern, ignoring it in practical deliberation. Strength of concern also does not amount simply to *intensity of feeling* at a given moment. I may care about the well-being of my friends more than I care about drinking soda, even though now, thirsty as I am, I experience my desire for soda considerably more intensely than I experience my concern for my friends. Yet intensity of feeling is not disconnected from depth of concern either.

Surges of emotion are not what constitute depth of concern, nor are they infallible indicators of it. Yet they strike me as one of three modestly reliable indices of concern, the other two being cognitive dispositions and motivational dispositions. Lacking a satisfactory consensus theory of desire, these signposts will have to suffice. As for emotions, the more one cares about morality, the more it colors one's emotional world, other things being equal (though other things are rarely equal when it comes to people's emotional lives). The morally concerned person tends to find the thought of doing wrong distressing—that is, she feels *guilt*. She also feels anger when reading about atrocities in the news, sadness when wondering, as Kant did, if "anything straight can be fashioned from the crooked timber of humanity," admiration for moral heroes, and so on. The *cognitive/perceptual* signs of moral concern are more subtle: a person concerned with morality is, to that degree, other things being equal, "morality conscious," noticing morally relevant things others might not. It is a feature of the human mind that we learn more about things of more concern to us: other things being equal, a bird lover will notice a bird on the roof while a person who does not care about birds might not. If one cares about morality, one is more apt to notice, for example, that a fellow human being is showing the signs of distress, or that a joke has the potential to offend certain people. Finally, *motivation* is an obvious guide to depth of concern. The more you care about something, the more it takes to stop you from acting for its sake (again, other things being equal). I shall focus here on motivational and emotional factors, but this is for convenience of discussion only.

---

[17] *A fortiori*, I shall not try to develop a fully-fledged positive view of what it means to be concerned about something in the first place. I do not, here, equate concern with caring as understood by Frankfurt, but I am not sure that I say anything here with which it is incompatible, given the assumption that wanting to have a desire does not entail thinking that one *should* have it.

Imagine a person who cares so much for her fellow human beings, or for what she rightly takes to be her moral duty to them, that she would act benevolently even if severe depression came upon her and made it hard for her to pay attention to others. Now imagine benevolence's fair-weather friend, who acts benevolently so long as no serious problems cloud her mind. Last, imagine the person who acts benevolently as a whim. It is Sunday morning and she is awakened by a call from a charity asking for a donation. Our agent thinks, "Oh, why not?" and is moved to do something right, so long as her credit card happens to be close enough to the bed. The first agent is more praiseworthy for her actions than the second, I hold, because of the first's greater moral concern, revealed by her persistent devotion to moral issues even in hard times. Kant, though, would have offered a different explanation, holding that the first agent is more praiseworthy than the second because she acts out of one motive, duty, while the second one acts out of a motive called "inclination," a basically hedonistic motive. This, however, need not be the case. My second agent, unlike Kant's own happy philanthropist, is not someone who "gets her kicks" out of doing good deeds (like Jane Austen's Emma in *Emma*), and her disposition need not be particularly sunny. She is just an ordinary person who does good for moral reasons, but whose moral concern is not deep enough to override some other concerns when they appear. Certainly, she sometimes feels a sense of satisfaction with herself, but one needs to be quite a psychological hedonist to think that any action that results in pleasure is motivated solely by it. There is no reason to say that her motives are different from the motives of the first agent, any more than there is a reason to say that the person who is truly devoted to her exercise program goes to the gym for different reasons than the person whose devotion is somewhat weaker. Both may go to the gym for health reasons, but one cares more about her health. Similarly, my first and second philanthropists both act benevolently for moral reasons, but one of them cares more. The contrast is *not* that happy philanthropists are less praiseworthy for their actions than sad philanthropists, but rather that *fair-weather*, frivolous philanthropists deserve less praise than those whose concern for morality or for the well-being of others is more *serious* or *deep*. The third agent—the person whose concern for morality is skin deep—may be called the *capricious* philanthropist, and should not expect much praise for an action that almost seems accidental, attributable to the charity's call and the location of the credit card more than to her depth of concern for her fellow human beings. Still, there is no reason to doubt that she acts for moral reasons. When a person whimsically asks for milk instead of cream in the coffee she has with her chocolate cake, one need not doubt that she does it for health reasons, one merely doubts the seriousness of her concern.

A natural question concerns Kant's misanthropic philanthropist who, in ordinary circumstances, must force himself to help people despite having no inclination to do so. Neither the misanthropic philanthropist nor the sorrowing philanthropist just discussed enjoys his action very much, which makes it deceptively easy to see them as similar instances of "acting out of duty." But each has a different story behind the fact that he drags his feet on his way to charity. The sorrowing philanthropist drags himself to action because he cares about the welfare of others so much that his concern moves him even when sorrow tempts him to stay home. The misanthropic philanthropist has to drag himself to his good works because his philanthropy, even at the happiest times, is *halfhearted*. This could mean that he cares about the good of others just enough to wish he cared more, or this could mean that his concern for morality is offset by essentially conflicting attitudes,[18] such as disdain for humankind. The fact that one's concern, whether for humanity or for one's wife or one's art, is enough to motivate one—albeit barely—even though one is grief-stricken, is a testimony to the strength of one's concern. The fact that one's concern, in the best of times, is barely enough to motivate one, shows a deficiency of concern or a half-heartedness. Thus, under many descriptions, the misanthropic philanthropist is less praiseworthy than he would be if it were not for his misanthropy. The picture becomes complicated, however, when one remembers how underdescribed Kant's case is. I have followed Hursthouse in taking the coldness of this philanthropist's heart to be a sign of halfheartedness. But truly, many different things can cause a person to appear cold or to experience himself as indifferent, and these things differ in morally significant ways (recall that what one cares about most is not always what one feels about most strongly and warmly). Also, while one's philanthropy may be halfhearted, it may be a part of a serious effort to habituate oneself into being a better person, a praiseworthy effort all by itself.[19]

---

[18] By 'essentially conflicting attitudes' I am referring to attitudes which conflict because of their content, and not merely because some accidental factor that makes it hard or impossible to act in accordance with both of them. For example, love of sports and love of philosophy may conflict accidentally if it so happens that one has to choose between a philosophy talk and a basketball game. The conflict here is a contingent one—it has to do not with the nature of sports or philosophy but rather with the fact that scheduling and the like do not permit one to divide one's time between the two. On the other hand, there is an essential conflict between one's love of philosophy and a craving to live the life of a "simple soul," between standard forms of Christianity and lust, and so on.

What about blameworthy actions? There are two types: those done for sinister reasons, and those done for morally neutral reasons as a result of some indifference to moral considerations. It stands to reason that a person who acts for sinister reasons is more blameworthy the stronger the ill will indicated by his action. Consider, however, a misdeed motivated by moral indifference. If philanthropists are judged by the depth of concern for relevant moral considerations that their actions express, wrongdoers can be judged by the depth of indifference to moral reasons that they reveal. The more moral concern required to take the right course of action, the less blameworthy an agent is for not doing so. Imagine that in a certain situation, it is right to act charitably and wrong to act uncharitably. Imagine also that, in this situation, it is much harder to act charitably if sad than if cheerful. The person who is charitable in sorrow shows herself to be more praiseworthy for her action than the person who is charitable generally, but not in sorrow. Likewise, it does not take a great deal of moral indifference to be uncharitable while sorrowing, but it does take a lot of such indifference to be uncharitable while happy. Thus, if the sorrowing philanthropist is more praiseworthy than the happy philanthropist, the sorrowing failed philanthropist is less blameworthy than the happy failed philanthropist. Hence, for example, the use of 'She was under stress' to excuse minor wrongs. In a more dramatic vein, if one wonders if actions such as hiding Jews from the Nazis are required by morality or are supererogatory, one compromise suggestion is that such actions *are* required; but performing them requires a degree of moral concern so rare that the person who fails to perform them shows hardly any measure of moral indifference, and thus is not particularly blameworthy.

---

[19] My discussion of the philanthropists owes a lot to Rosalind Hursthouse, "Ethics and the Emotions," in Daniel Statman, ed., *Virtue Ethics* (Washington, DC: Georgetown, 1997), pp. 99-118; and Philippa Foot, *Virtues and Vices* (Berkeley: California UP, 1978).

If depth of one's concern for the right-making features of one's action changes an action's moral worth, how does Huckleberry measure up? He merits praise for helping Jim, because he is acting for the right reasons—he performs his action because of its right-making features. But can he be said to be acting out of deep concern for the right-making features of his action? Yes—at least in the scenario imagined above. Recall that concern does not amount to reflective endorsement and does not have to be conscious. Let us compare Huckleberry to another inverse akratic, whose action I take not to be morally praiseworthy. Joseph Göbbels, as is evident from his diaries and some of his public speeches, suffered from surprisingly frequent attacks of what he called "weakness of will," which he attributed to fatigue and stress, and which consisted in feeling compassion for the victims of the Nazi regime. Göbbels repressed and overcame them with relative ease, the way a civil person may overcome a desire to be rude at a family dinner. Let us suppose that on a certain occasion, Göbbels, tired at the end of a long day, finds himself momentarily unable to resist his compassion. Against his best judgment, he makes a low-risk compassionate gesture, such as permitting a Jewish acquaintance to leave Germany. Assume he acts for the same reasons as Huckleberry. Still, his action does not impress us as Huckleberry's does. Göbbels's inverse akratic compassion strikes us as shallow or capricious, not so much because of the short time his attacks last, but because of the ease with which they are overcome. His compassion is also offset by the strength of his essentially conflicting attitudes—his deep racism and so on. His action seems the product of fatigue or stress more than anything else. We can imagine his friends using this as an excusing condition from blame, and for us it is "excusing" from praise. Things appear to be different with Huckleberry. He does not need fatigue or alcohol to help Jim. As he is described in the novel, one gets the impression that he helps Jim because his concern for his fellow human beings in general, and for Jim as a particular human being, is strong enough to make him act despite reservations. The strength of his concern for Jim as a person is a testimony to the shallow nature of his racist attitudes, not the other way around. Thus, even though both Huckleberry and Göbbels act—imagine—for the right reasons, and both of them act against their best judgments, we view them differently. One of them appears to be a good (though imperfect) boy who performs a morally praiseworthy action, while the other appears to be a bad man who, thanks to fatigue or stress, performs a morally desirable but negligibly praiseworthy action.

VI. CHARACTER?

It is tempting to rephrase what I have said about Huckleberry in terms of character. His human concern seems to be a deep feature of his character, while his racist convictions seem not to be. It is no accident that talk of character comes naturally here. Just as the idea of doing the right things for the right reasons, or out of concern for the right-making features of one's action, accounts for the appeal of the doctrine of the motive of duty, the idea that deep concern for these features is worth more than shallow concern for these features accounts for the appeal of the Aristotelian idea that right (or fine) actions are only praiseworthy (or virtuous) if they follow from the agent's *character.*

Consider Aristotle's treatment of bravery. For an instance of defending one's city in war to be praiseworthy, it is not enough for Aristotle that it is performed for the right reasons. However pure his motive is, if he is capable of defending his city only because he is pathologically fearless, drunk, or such a skilled soldier that war does not frighten him, an agent's action will still not be morally praiseworthy. If, like Kant's misanthropic philanthropist, he needs the aid of soldierly self-control to drag himself to action, he is only partially virtuous. The *fully* morally praiseworthy agent is the man who defends his city because he fears moral disgrace more than he fears death: he cares more for what is right than for his life. Such deep concern for the right-making features of defending one's city is to be found in the *brave* person—the person in whom risking his life for the sake of doing the right thing expresses a virtue of character. The virtuous person is different from others in possessing the markers of moral concern. He has a *die-hard disposition* to do what is fine (defend the city): he has a different *emotional* life from that of the nonvirtuous person (the thought of moral failing disturbs him) and he has an ability to *recognize* the right occasions for certain kinds of action which a mere theoretical knowledge of ethics does not give us.[20]

Why should Aristotle, or anyone else, believe that the praiseworthiness of an individual action depends on the character from which it stems? If one thinks of character as a stable disposition of some sort, the idea may seem strange. Why should the moral worth of the action that Steve performs on November 13, 1999, be any different if he has performed similar actions for years and will most likely continue to do so?

---

[20] For a discussion of related themes, see John McDowell, "Virtue and Reason," *The Monist*, LXII (1979): 331-50.

The answer is that the mere frequency or predictability of an action does not matter at all to the moral worth of the actor, but these things may be signs of something relevant: *deep moral concern*. The pathologically fearless man or the well-trained soldier may have just as stable a disposition as the brave man to defend his city, but fearless or merely well-drilled actions do not express courage.[21] Consider Steve, a teacher, who has not missed class in twenty-six years. This says something good about him precisely because it is a sign of his devotion to his students and his job. Such evidence, however, can be overridden. Imagine, for example, that Steve's disposition to is due to a blind adherence to a work ethic which would be manifested even if he were the manager of children working in an unsafe fireworks factory. If so, we would no longer be inclined to think that Steve's disposition says anything particularly good about him. His devotion to his students is his virtue: his actions over the last twenty-six years are mere evidence—incomplete, defeasible evidence—of this virtue. If we wish to be neo-Aristotelian, we might want to say that Steve's coming to class on November 13 despite his severe migraine is especially praiseworthy because it stems from a virtue of character. This is untenable if "stems from a virtue of character" means only "stems out of a stable disposition," but makes perfect sense if it means "stems out of a markedly deep morally relevant concern"—in this case, deep devotion to his students. Note that "deep concern" does not automatically mean "long-lived concern." It happens to be empirically true that deep concerns do not generally change overnight: Dr. Jekyll is science fiction. On the rare occasions when deep concerns do undergo rapid change, the resulting actions are not capricious, but on a par with those produced from long-standing concern. For example, if it were known to us that that a teacher, previously indifferent to his students, underwent a conversion of sorts, after having read Bel Kaufman's *Up the Down Staircase*, and developed a genuinely deep concern for his students, we would have no reason to regard his devotion to his students as any shallower than Steve's, even if the conversion happened only a few months ago. We might say that the person's character changed. Of course, nobody knows for sure if such a person's new concern is deep or not, and so we tend to postpone our judgments until we see if it lasts; but the "test of time" is only that—a mere *test* for the depth of a concern, and an imperfect one at that. Longevity does not make a concern deep any more than its lack makes a concern shallow—in a comfortable climate, one may be a fair-weather friend for many years. Thus, the idea that character matters for the moral worth of individual actions makes sense if one takes character to be about depth of concern and not about predictability or frequency. Many traditional objections to character-oriented views lose their force in this way. To give only one example, suppose that as the Nazis come to power, your long-time fair-weather friend, who has never done you wrong before, cheerfully informs on you. A traditional objection states that virtue ethics is committed to excusing his action as "out of character." But if "in character" does not mean "predictable," or "in keeping with historical trends" then "unpredictable" and "out of keeping with trends" does not always mean "out of character," either. In defense of the critics of virtue ethics, however, it can be said that virtue ethicists themselves have focused on predictability and stability, to their own detriment.

---

[21] The connection between virtue and depth of concern is evident, though never quite explicit, in Foot.

Unfortunately for traditional virtue ethicists, even stable, deep concern for morality (or concern for anything else) results in little predictability or stability of behavior, as has been shown by various psychological experiments[22] but as can also be deduced from more intuitive evidence. A corrupt lawyer may be more likely to return your book on time than an activist whose life is devoted to global justice, and soldiers who are brave in battle often fear public speaking. Likewise, one cannot infer the degree to which busy university faculty value friendship or conversation from their willingness to chat in the halls: some grouchy professors are merely nearing tenure decisions, some talkative professors simply have more time as a result of being on leave, and so on. Even a talkative person nearing a tenure decision is not certain to be displaying her strong desire for affiliation: a childhood in which strict norms of politeness were imposed upon her might be a much stronger factor. In a world in which people are rarely allowed to act on their hearts' desires, one's actions do not always reflect one's concerns very well.

While these tales show problems for the traditional notion of a character trait, they do not show that there is no fact of the matter about moral concern. The person who loves more is not always the person who says "I love you" more often, but some people do love more than others. The person who comes to class more often is not always the person who is more devoted to his students, but some people are more devoted to their students. And the connection between depth of concern and moral worth is an important truth aimed at by talk of character.

---

[22] See John Doris, *Lack of Character Personality and Moral Behavior* (New York: Cambridge, forthcoming); and Gilbert Harman, "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error," *Proceedings of the Aristotelian Society*, xcix (1998-99): 315-31.

VII. A FEW MORE CLARIFICATIONS

This account of moral worth raises a tremendous number of questions. In this final section, I attempt to answer a few of the most pressing ones concerning self-control, misguided conscience, and moral knowledge.

First, is there not some merit in self-control, which characters like Huckleberry lack? The fact that he helps Jim despite his convictions might show that Huckleberry has very little self-control; or it may show that his concern for Jim's humanity is so strong that it overwhelms normal self-control, and I was assuming the latter, which would make Huckleberry paradoxically similar to Aristotle's brave person. The brave person's desire to avoid disgrace is stronger than his fear of death. Huckleberry's well-motivated desire to help Jim overrides his attachment to his misguided convictions. The brave person's desire to defend his city is strong enough that it does not need the aid of self-control to win his inner struggle. Similarly, Huckleberry's concern for Jim's humanity is strong enough that it does not need the aid of unusually weak self-control to win the inner struggle. In the cases of both Huckleberry and the brave person, it is a virtuous concern that wins the struggle, and that is all that matters to the moral worth of their actions.

What of misguided consciences?[23] Can there be anything praiseworthy about a person's devotion to a morally wrong cause? The conclusion that seems to follow from my view is that there cannot be. Moral praiseworthiness is rooted in responsiveness to what in fact are moral reasons, not to whatever it is that one takes to be moral reasons, and so the fact that the advocate of the bad cause thinks she is acting for moral reasons does not by itself redeem her. The reasons to which the advocate of the bad cause responds are in fact antimoral reasons, in which case she acts malevolently, or at best reasons which accidentally conflict with morality, in which case she shows moral indifference. How can there be anything good in this?

There could be none, if it were true that being an advocate of an ideology entailed having the beliefs required by the ideology, and if being an advocate of a person or a country entailed being an advocate of what the person or the country really are. Things are much more complicated than this, however. Klaus Mann provides us in *Mephisto*[24] with the character of Hans Miklas, a young, uneducated, naïve man who becomes a member of the Nazi party while it is still a fringe group. Yet the reader gets the impression that Miklas is not really a bad man. His attraction to Nazism appears to rely a lot on his desire for social justice and his disgust with corruption. In his mind, he associates "German honor," not with conquering other countries, but with a public life free of corruption. After Adolf Hitler comes to power, Miklas, despite being promoted and respected due to his long-standing party membership, quickly discovers that not only has social injustice remained ubiquitous, but it has in fact increased, and so has corruption. He is so devastated that he consciously destroys himself by publicly denouncing the regime. This action authenticates the suspicion that, from the start, his attraction to Nazism was a morally mixed bag, containing an element of true attraction to justice. Can one really act for a mix of morally good and bad reasons? It seems Miklas does, because some frustration and hatred have obviously affected his infatuation with the Nazi movement instead of some other workers' movement and made him all too willing to believe anti-Semitic propaganda—but a true concern for justice surely contributes to his great disappointment with the Nazi regime. (To complicate the character even further, Miklas's Nazism is also the result of his ignorance and cognitive limitations, which makes it hard to tell innocent false beliefs from irrational, motivated prejudice.) It may seem, from a Kantian[25] point of view, that one either is motivated by morally relevant reasons or is not, without the possibility of mixed motives, but this is not the way we think when we judge people's motivations. Granted, in some contexts, it may seem as if it is. To borrow an observation from Christine Korsgaard,[26] if a student tells us that he has decided to study calculus for the sake of the intellectual challenges involved, but we learn that it is also a required class, we tend to take a cynical attitude toward his claim. But consider a student who had always wanted to take calculus but been afraid of doing so because she doubted her abilities. Upon entering college, she discovers a degree requirement can be met by one of a few classes, one of which is calculus. She decides to take calculus, thankful for the requirement at the same time as she fears it. While this student would not have taken calculus if it were not required, it would be unjust on the part of her math professor to treat her as on par with her peer who is dragged, kicking and screaming, into the required class. Moral motives admit of the same complications.

---

[23] Hill, "Four Conceptions of Conscience," *Nomos*, LX (1998): 13-52.

---

[24] Robin Smyth, trans. (New York: Random House, 1977).

[25] See Herman.

[26] *The Sources of Normativity* (New York: Cambridge, 1996), p. 105.

Finally, what of the view that there is something especially good in having knowledge of the virtues, correct moral convictions, and the ability to deliberate well on moral matters? If acting for moral reasons is not the same as acting for what we take to be moral reasons, and depth of moral concern is not the same as conscious commitment to moral principles, how can this be true? I have conceded earlier that Huckleberry would be a better person if his conscious moral convictions were not appalling. But some will argue that, in addition to this fact, there is something especially regrettable in the fact that Huckleberry's character deficiency is located specifically in his ability for moral *deliberation*, as opposed to somewhere else. I believe that there is something to this claim, but that the goodness of having the ability to deliberate well about morality can be maintained without denying that Huckleberry's action is morally praiseworthy. First of all, an agent who does not know his virtues from his vices is likely to try to make himself a worse person, whereas the person who knows his virtues from his vices is likely to try to make himself a better person. More importantly, there are certain types of morally desirable actions that are very hard to perform if one does not have the right moral principles, or at least the ability to deliberate well on moral matters. For example, it is very hard to vote for the right political candidate if one cannot deliberate well on moral (and other) matters. While some of those who voted for Hitler were obviously bad people, others strike us as morally average people (or, in the case of Miklas, somewhat better) whose unreflectiveness, simplicity, and incompetence at deliberation made them easy prey to a candidate who promised jobs, national pride, and order in the streets. Given the complex moral decisions that even the most mundane life has to offer, it seems that we have a moral duty not to allow ourselves or our children to be too stupid, unreflective, or uninformed, especially with regard to morally relevant issues, and thus the fact that a person—even a good person—cannot deliberate very well or is ignorant when it comes to morally relevant matters has a special sadness to it.

It has been my goal here to sketch a quality-of-will based theory of moral worth, and show that such a view is potentially useful in understanding some facts of moral life. On my view, people are praiseworthy for acts of good will and blameworthy for acts of ill will or the absence of good will, and the amount of praise or blame they deserve varies with the depth of their motivation or the extent of their indifference. But since, I hold, good will is wanting to perform actions that have whatever property it is that makes actions right, a full account of moral worth is impossible until we know what property it is that makes actions right. Other gaps remain in my account, such as the significance of my view for moral responsibility and autonomy.[27] I leave these topics for another occasion.

NOMY ARPALY

Rice University

---

[27] As a view of blameworthiness and praiseworthiness which does not rely on the concept of freedom my view is a potential basis for a compatibilist theory of moral responsibility. I discuss moral responsibility and autonomy in the fourth and fifth chapters of my forthcoming book *Unprincipled Virtue: An Inquiry into Moral Agency* (New York: Oxford, 2002). I plan to give an extended treatment of free will per se in future work.

WHAT IS THE PROBLEM OF COHERENCE AND TRUTH?[*]

In inquiry, we often proceed by gathering information from different sources that may not be very trustworthy in themselves, and all we have to go on may well be the degree to which the reports cohere. If they do cohere or agree to a large extent, we tend to think that the information is credible. For instance, if the first dubious witness to be queried says that John was at the crime scene, the second that John has a gun, and the third that John transferred a large sum to his bank account shortly after the robbery, then the striking coherence of the different testimonies (their individual dubiousness notwithstanding) would normally make us pretty confident that John is to be held responsible for the act.

Impressed with this sort of example, some philosophers have concluded that coherence is the one single thing separating *warranted* from *unwarranted* beliefs. Roughly speaking, beliefs that hang well together, agree, or exhibit mutual support are (according to philosophers of this inclination) thereby justifiably held and, if true, known to be so. Others have been more modest in their philosophical extrapolations from such pieces of commonsense reasoning. For C. I. Lewis,[1] for example, coherence cannot create credibility from scratch; it can only amplify an already existing positive degree thereof. Yet coherence has an important role to play also in Lewis's foundationalist framework where it is needed for building up probabilities "sufficient for rational and practical reliance" (AKV, 357).

Lewis, as we just saw, can be interpreted as suggesting a probabilistic rendering of what it means for coherence to imply truth. But what exactly is the role of coherence, and what are the probabilistic

---

[*] Thanks are due to Ludwig Fahrbach for patiently going through several earlier versions of this article with me. I have also benefited from the criticism and suggestions of Hans Rott and Tomoji Shogenji. My research was financed by the DFG (*Deutsche Forschungsgemeinschaft*) as a contribution to the project, *Logik in der Philosophie*.

[1] *An Analysis of Knowledge and Valuation* (La Salle: Open Court, 1946); hereafter AKV. Lewis sees himself rather as advocating what he calls "conceptual pragmatism," a philosophical position situated "somewhere between [William] James and the absolute pragmatism of [Josiah] Royce" (in his "Autobiography," in P. A. Schilpp, ed., *The Philosophy of C. I. Lewis* (La Salle: Open Court, 1968), p. 17).

0022-362X/02/9905/246–72

facts here? In their search for answers, a number of recent authors[2] have focused on the problem of whether more coherence implies a higher likelihood of truth. Let us say that coherence is *truth conducive* if it has that property. The question then is: If system *S* is more coherent than system *S'*, *are we then allowed to conclude that* *S* is more likely than *S'* to be true as a whole?

There are good reasons to pay attention to this particular question. First, as Peter Klein and Ted A. Warfield[3] point out, it asks for a minimal sense in which coherence could imply truth. It would seem difficult to maintain that coherence implies truth without also maintaining that more coherence implies a higher likelihood of truth. Second, it is relatively clear and unambiguous. By contrast, the question 'Is a coherent system highly likely to be true?' suffers from serious vagueness—indeed, along two dimensions: how high a degree of coherence it takes for a system to qualify as "coherent," and how high a degree of likelihood it takes for a system to qualify as "highly" likely to be true.

In the next two sections, the task will be, first, to get clearer on what kind of property coherence is, and, second, to obtain a better understanding of how truth conduciveness should be more precisely construed. It is implausible to think that coherence is truth conducive in the absence of further conditions: a well-composed novel is usually not true, and yet it may still be highly coherent—perhaps far more so than reality itself. This raises the question of what the additional prerequisites might be, a topic that will be dealt with in sections III and IV. In the final section, I shall return to the presystematic question of whether coherence implies truth and consider a more ambitious rendering of it.

One of my theses here will be that common criticisms against a connection between coherence and truth are ill-founded, because they rest on an inadequate and uncharitable understanding of truth conduciveness. But I shall also argue that even on a more adequate rendering of that notion, coherence is at best truth conducive in a very weak sense.

---

[2] See Peter Klein and Ted A. Warfield, "What Price Coherence?" *Analysis*, LIV, 3 (July 1994): 129-32, and "No Help For the Coherentist," *Analysis*, LVI, 2 (April 1996): 118-21; Trenton Merricks, "On Behalf of the Coherentist," *Analysis*, LV, 4 (October 1995): 306-09; Charles B. Cross, "Coherence and Truth Conducive Justification," *Analysis*, LIX, 3 (July 1999): 186-93; and Tomoji Shogenji, "Is Coherence Truth-Conducive?" *Analysis*, LIX, 4 (October 1999): 338-45. See also Luc Bovens and Olsson, "Coherentism, Reliability and Bayesian Networks," *Mind*, CIX (October 2000): 685-719, and my "Why Coherence Is Not Truth-Conducive," *Analysis*, LXI, 3 (July 2001): 236-41.

[3] "What Price Coherence?" p. 129.

I. THE CONCEPT OF COHERENCE

Lewis defined coherence—or 'congruence',[4] to use his favored term—as follows:

> A set of statements, or a set of supposed facts asserted, will be said to be congruent if and only if they are so related that the antecedent probability of any one of them will be increased if the remainder of the set can be assumed as given premises (AKV, 338).

The set S consisting of propositions $A_1,\ldots A_n$ is congruent relative to a probability distribution P just in case $P(A_i|B_i) > P(A_i)$ for $i = 1,\ldots,n$, where B_i is a conjunction of all elements of S except A_i.

There are two reasons why I shall choose not to follow Lewis in this respect. The first is trivial: since my concern is with the question whether more coherence implies a higher likelihood of truth, I am interested, first and foremost, in a notion of coherence admitting of (nontrivial) degrees, whereas Lewis is here proposing an absolute (nongraded) conception.[5]

The second reason is more fundamental. As the following example shows, congruence fails even as a reasonable explication of absolute coherence. The reason why it fails is of relevance to my concerns.[6] Suppose that there is a reasonable number of students and a reasonable number of octogenarians. Suppose that all and only students like to party and that all and only octogenarians are bird watchers, and that there are some, but very few, octogenarian students. A murder happened in town. Consider the following propositions:

A_1 = 'The suspect is a student'.

A_2 = 'The suspect likes to party'.

A_3 = 'The suspect is an octogenarian'.

A_4 = 'The suspect likes to watch birds'.

The set $S = \{A_1,A_2,A_3,A_4\}$ is congruent in Lewis's sense, yet intuitively anything but coherent: one half of the story (the one about the partying student) is very unlikely given the other half (the one about the bird-watching octogenarian).

---

[4] Lewis prefers 'congruence' to 'coherence' because he wants to mark his departure from British post-Kantian idealism and its "coherence theory of truth."

[5] Lewis extrapolates a notion of "extent of congruity" (AKV, 357) from his absolute notion of congruence but leaves the details of this nontrivial undertaking in the dark.

[6] The following counterexample is adopted from Bovens and Olsson.

What has gone wrong here? One notable thing is that the joint probability of the propositions in S is very low; the probability that the suspect is both a partying student and an octogenarian bird watcher is close to zero. This suggests identifying the degree of coherence of a set with its joint probability.

$$C_0(A,B) = P(A \wedge B)$$

This measure takes on a minimum value of 0 if and only if there is no overlap between A and B. It takes on a maximum value of 1 just in case $P(A) = P(B) = 1$, that is, just in case both A and B are for sure. It is obvious how to generalize this measure to the case of an arbitrary (finite) number of propositions:[7]

$$C_0(A_1,\ldots,A_n) = P(A_1 \wedge \ldots \wedge A_n)$$

Suppose, however, that there has been a robbery. To get an unbiased view on who might have committed the crime, you decide to consult four different witnesses. Suppose that the first two witnesses both claim that Steve did it, the third witness that Steve, Martin, or David did it, and the fourth witness that Steve, John, or James did it. Which pair of statements is the more coherent: that delivered by the first two sources or that delivered by the last two sources? It is difficult to escape the feeling that the statements delivered by the first two sources are more coherent in an intuitive sense. After all, unlike the last two witnesses, the first two say exactly the same thing, so that the degree of agreement is higher. According to the C_0 measure, by contrast, the degree of coherence is the same, equaling the probability that Steve did it.

This leads us directly to the next proposal. Rather than measure the overlap of propositions, we might measure the extent to which they agree. The more the propositions agree, the more coherent they are. From this perspective, propositions that coincide are always maximally coherent, even if their joint probability is not very high. A simple way to measure the extent of agreement is the following:

$$C_1(A,B) = \frac{P(A \wedge B)}{P(A \vee B)}$$

$C_1(A,B)$ measures how much of the total probability mass assigned to either A or B falls into their intersection. $C_1(A,B)$ takes on values between 0 and 1. As before, the degree of coherence is 0 if and only if $P(A \wedge B) = 0$, that is, just in case A and B do not overlap at all, while the degree of coherence equals 1 if and only if $P(A \wedge B) = P(A \vee B)$, that is, just in case A and B coincide. The measure is straightforwardly generalizable:

---

[7] I shall be concerned throughout exclusively with the coherence and truth of finite sets.

$$C_1(A_1, \ldots, A_n) = \frac{P(A_1 \wedge \ldots \wedge A_n)}{P(A_1 \vee \ldots \vee A_n)}$$

An alternative measure of coherence was recently introduced by Tomoji Shogenji (*op. cit.*):

$$C_2(A,B) = \frac{P(A|B)}{P(A)} = \frac{P(A \wedge B)}{P(A) \times P(B)}$$

Like the other measures, $C_2(A,B)$ equals 0 if and only if A and B do not overlap. But while C_1 takes on its maximum value when the propositions coincide regardless of how specific those propositions are, this is not so for C_2. For suppose that A and B coincide. Letting $x = P(A)$ we get $C_2(A,B) = P(A \wedge B)/P(A) \times P(B) = x/x^2 = 1/x$. Hence, the lower the probability of the coinciding propositions is—that is, the more specific the information is—the higher is the coherence.[8] This measure, too, allows for effortless extension to the general case:

$$C_2(A_1, \ldots, A_n) = \frac{P(A_1 \wedge \ldots \wedge A_n)}{P(A_1) \times \ldots \times P(A_n)}$$

The difference between C_1 and C_2 stands out most clearly when two pairs of coinciding propositions are contrasted which differ with respect to specificity. Let me modify the robbery example somewhat. Suppose, as before, that the first two witnesses report that Steve did it, and that the third witness reports that Steve, Martin, or David did it. But suppose now that the report of the fourth witness coincides with that of the third, so that they both report that Steve, Martin, or David did it. Which testimonies are now the most coherent, those of the first two witnesses or those of the last two? Both sets are in full agreement, and so their degree of coherence should presumably be the same. Appealing to our intuitions of mutual support yields the same result; for each set, the one statement in the set is established if the other is assumed as given premise. This is also what we get if we apply C_1. On the other hand, the agreement between the first two testimonies is surely much more striking since, unlike the last two, they coincide on a very specific statement. Therefore, one could argue, the degree of coherence should also be greater in the first case. This is also what C_2 yields.

---

[8] C_2, of course, is nothing but one of Rudolf Carnap's confirmation measures discussed at length in his *Logical Foundations of Probability* (Chicago: University Press, 1950), chapter 5. The other Carnapian confirmation measure, $C(A,B) = P(A|B) - P(A)$, cannot be used in this context, as it is not symmetric in its arguments.

There may be many other ways of measuring the degree of coherence or agreement. While both C_1 and C_2 have some initial appeal, I doubt that our intuitions are clear enough to single them out as the only plausible candidates. Nonetheless, they do provide us with a useful starting point for concrete discussion.

II. PROPOSITIONAL VERSUS DOXASTIC TRUTH CONDUCIVENESS

It appears that most philosophers[9] who have discussed the coherence and likelihood of truth of sets of propositions have thought of the relevant likelihood as one that the whole set be true, that is, the likelihood of *joint* truth. In other words, the relevant probability is the probability of the *conjunction* of all propositions in the relevant set.

This leads me to my first attempt to explicate what it means for coherence to be truth conducive. A measure of coherence will be said to be propositionally truth conducive if and only if a higher degree of coherence implies a higher probability of joint truth.

Definition 1: a measure C of coherence is *propositionally truth conducive* if and only if the following holds: if $C(A_1,\ldots,A_n) > C(B_1,\ldots,B_m)$, then $P(A_1\wedge\ldots\wedge A_n) > P(B_1\wedge\ldots\wedge B_m)$.

Note that n and m might be distinct; that is, the size of the sets is allowed to vary. The size issue will be addressed at greater length in section IV.

Are there any (nontrivial) propositionally truth conducive coherence measures?[10] A measure is truth conducive, in this sense, if more coherence means higher joint probability. According to C_0, coherence *is* the joint probability, and so this measure, unsurprisingly, comes out as truth conducive in the propositional sense. This fact actually tells against, rather than in favor of, the reasonableness of our explication. C_0 is not a plausible measure of coherence, and it would be highly surprising, therefore, if it turned out to be, in any interesting sense, truth conducive.

---

[9] For two exceptions, see Merricks; and, following him, Shogenji, especially p. 344.

[10] Measures that do not make discriminations as for coherence are trivially truth conducive since the antecedent of the implication in definition 1 will always be false. Such measures are disregarded in the following.

More evidence points in the same direction. Both C_1 and C_2, two initially attractive measures, turn out not to be truth conducive in the sense of definition 1. The (unmodified) robbery example in section I provides us with a counterexample to the propositional truth conduciveness of C_1. Recall that the first two witnesses claimed Steve to be the culprit, the third that Steve, Martin, or David did it, and, finally, the fourth that Steve, John, or James is to blame. As we noted before, the first two statements are more C_1-coherent than the latter two, and yet the joint probability of the two sets is the same. Concerning C_2, it was shown in section I that among pairs of coinciding propositions it yields a higher coherence value for more specific pairs. On the other hand, more specific coinciding propositions have a lower (joint) probability.[11]

What conclusions can be drawn from these difficulties? One possible reaction is to conclude that they teach us an important lesson about the connection between coherence and truth, or rather about the lack of such a connection. This is the path taken by Klein and Warfield. Appealing to a detective story and to our intuitions about coherence in that particular case, they claim to have established that coherence is not truth conducive in the above sense.[12] Taking the theory of Laurence Bonjour[13] to be a case in point, they conclude that coherence theories of knowledge that rely on coherence to be truth conduciveness are profoundly mistaken.

Another possible response (which I have already hinted at) is to question the explication of truth conduciveness. Following this track, Shogenji notes (*op. cit.*), as I have also done, that if coherence is measured using the measure here called C_2, then a stronger (that is, more specific) set may well be more coherent, though it is less likely jointly to be true, so that C_2-coherence does not come out as truth conducive (in the above sense). Shogenji now suggests that this conception of truth conduciveness is mistaken. By the ***total individual strength*** of a set of propositions I mean the product of the probabilities of the individual propositions in the set. Thus, the total individual strength of $\{A_1,\ldots,A_n\}$ is $P(A_1)\times\ldots\times P(A_n)$. The claim is that when assessing the truth conduciveness of coherence "we need to check whether more coherent beliefs are more likely to be true together than less coherent *but individually just as strong beliefs*" (*op. cit.*, p. 342); that is, the only relevant comparisons, on this view, are those among sets having the same total individual strength. What Shogenji proposes, then, is the following: a coherence measure C is truth conducive if and only if: if $C(A_1,\ldots,A_n) > C(B_1,\ldots,B_m)$ and $P(A_1)\times\ldots\times P(A_n) = P(B_1)\times\ldots\times P(B_m)$, then $P(A_1\wedge\ldots\wedge A_n) > P(B_1\wedge\ldots\wedge B_m)$. He observes that on this test, which he thinks is "more reasonable" (*op. cit.*, p. 342), C_2-coherence is truth conducive.[14] I have argued elsewhere[15] that Shogenji does not have a good case for fixing the total individual strength, a point to which I shall return in section IV.

---

[11] These counterexamples also disprove the weaker claim that C_1 and C_2 are propositionally truth conducive among sets of propositions of the same size.

[12] See their "What Price Coherence?" The detective story runs as follows (pp. 130-31). A detective has gathered a large body of evidence that provides a good basis for pinning a murder on Mr. Dunnit. In particular, the detective believes that Dunnit had a motive for the murder and that several credible witnesses claim to have seen him do it. But because the detective also believes that a credible witness claims that she saw Dunnit two hundred miles away from the crime scene at the time the murder was committed, her belief set is incoherent (or at least somewhat incoherent). Upon further checking, the detective discovers some good evidence that Dunnit has an identical twin whom the witness providing the alibi mistook for him. Now the extended set is formed by adding the belief about the twin, which is not implied by what was previously believed. It follows by probability calculus that the extended set is less likely to be true than the original, and yet the former is arguably more coherent than the latter. Klein and Warfield conclude that "coherence, per se, is not truth-conducive" (p. 131).

[13] *The Structure of Empirical Knowledge* (Cambridge: Harvard, 1985).

There is also the option of blaming the troubles on our preliminary measures of coherence. But whether or not one ultimately finds these measures unsatisfactory, there is a fundamental reason to be dissatisfied with our explication of truth conduciveness, a reason different from the unconvincing one offered by Shogenji. The crucial observation is that what we are primarily interested in, from an epistemological perspective, is the coherence and truth of *beliefs*, not the coherence and truth of sets of propositions of a believer considered in the abstract.[16] In other words, the interesting question is not whether more coherent propositions are more likely jointly to be true in general, but whether this is so among propositions that are actually held as beliefs.

---

[14] Shogenji's proposal to fix the strength is made in the context of his criticism of Klein and Warfield's view and does not represent his own approach to the evaluation of truth conduciveness. The position that he finally arrives at is that truth conduciveness should not be evaluated at the level of sets but at the level of individual beliefs. This leads him to conclude that coherence is *not* truth conducive after all because, he argues, "coherence per se has no bearing on the truth of individual beliefs" (*op. cit.*, p 344).

[15] See my "Why Coherence Is Not Truth-Conducive."

[16] Hilary Putnam makes a similar observation in his *Pragmatism. An Open Question* (Cambridge: Blackwell, 1995): "[c]oherence theorists have always pointed out that what they require for truth is not mere coherence of *sentences* but coherence of *beliefs*" (p. 64).

Presumably, Klein and Warfield would agree that the important question is what holds for believed propositions. For they declare that they want to show "that by increasing the coherence of a set of *beliefs*, the new, more coherent set of *beliefs* is often less likely to be true than the original, less coherent set."[17] Unfortunately, they lose sight of the distinction between propositions in general and believed propositions in their probabilistic argumentation. By the same token, Shogenji is officially concerned with beliefs, unofficially with bare propositions.[18]

Let me state the alternative proposal with more precision. We let $\mathrm{Bel}_S A$ stand for 'The subject S believes that A'.

Definition 2: a coherence measure C is *doxastically truth conducive* (for S) if and only if: if $C(A_1,...,A_n) > C(B_1,...,B_m)$, then $P(A_1 \wedge ... \wedge A_n | \mathrm{Bel}_S A_1,...,\mathrm{Bel}_S A_n) > P(B_1 \wedge ... \wedge B_m | \mathrm{Bel}_S B_1,...,\mathrm{Bel}_S B_m)$.

The definition says that a measure of coherence is doxastically truth conducive just in case a more coherent set of believed propositions is more likely to be true than a less coherent set of believed propositions. The subject S will usually be omitted. I shall refer to Bel$A$ as a *belief report* and to A as the *content* of that report. The difference between the two conceptions of truth conduciveness is that the doxastic conception, unlike its propositional counterpart, is a conditional notion: it conditionalizes on the assumption that the propositions in question are believed (by a given subject). In the next section, I shall put forward some conditions that seem required for coherence to be doxastically truth conducive. The remainder of this section is devoted to the elucidation of definition 2.

We could get this far without touching the difficult subject of how to interpret probability statements. But whether definition 2 makes good sense does depend, to a nonnegligible degree, on how probability is construed. Suppose that we decide to adopt a subjective interpretation. On one influential view, fully believing a proposition requires assigning a subjective (personal, credal) probability of 1 to that proposition.[19] Therefore, the probability of $A_1 \wedge ... \wedge A_n$, given that I believe each conjunct, would also be 1. But then the consequent of the defining condition in definition 2 expresses the falsity that $1 > 1$, which was obviously unintended and leads to trivial results. Indeed, on a subjective rendering of probability coherence itself is an entirely trivial matter if it is applied to propositions already believed: the coherence of any set of beliefs is 1, regardless of whether coherence is measured by C_1 or C_2, or, it would seem, any other reasonable measure.

---

[17] "What Price Coherence?" p. 129, my emphasis.

[18] Shogenji addresses coherence of believed propositions in "The Role of Coherence in Epistemic Justification," *Australasian Journal of Philosophy*, LXXIX, 1 (March 2001): 90-106.

[19] For a prominent example, see Isaac Levi, *The Fixation of Belief and Its Undoing* (New York: Cambridge, 1991).

Fortunately, there are independent reasons not to interpret probability subjectively in this context. In order to retain the connection to the original, presystematic problem of coherence and (objective) truth, it is customary even among those who advance a propositional conception of truth conduciveness to adopt an objectivist interpretation of probability.[20] I shall do so, too. In leaving it open what objective probability means more precisely, I also follow what seems to be the mainstream. The conclusion here will eventually be that coherence is at best truth conducive in a weak sense, even if we objectify coherence. It goes without saying that the connection between coherence and truth, if there is such a connection at all, will be no stronger under other, nonobjective interpretations. Thus, from the point of view of coherence and truth, we did the coherence theorist a favor when we objectified coherence.[21]

The simple observations that led us to conclude that coherence, in the sense of C_1 or C_2, is not propositionally truth conducive do not carry over to the doxastic case. When refuting the propositional truth conduciveness of C_2, for instance, we referred to cases of two pairs of coinciding propositions, one pair more specific than the other. We noted, on the one hand, that C_2 yields a higher coherence value for the more specific case and, on the other hand, that the (joint) probability will in this case be lower. By probability is here meant the *antecedent* joint probability. But from the fact that the antecedent joint probability is lower for the more coherent case, it does not follow that the *posterior* joint probability—the joint probability conditional on belief—is also lower.

Thus, the move from propositional to doxastic truth conduciveness blocks the general inference from a lower antecedent joint probability to a lower likelihood of truth. This notwithstanding, there are cases in which a lower joint probability, and a correspondingly higher degree of C_2-coherence, is in fact accompanied by a lower likelihood of truth (in the conditional sense). Compare agreement on two tautologies with agreement on two more specific propositions. The C_2-coherence will be higher in the more specific case, and yet the joint probability in the case of the tautologies will be 1, regardless of whether we are talking about conditional or unconditional probability, and so there is no room for the probability to be higher in the more coherent scenario.[22] The general lesson to be learnt from this observation is that a coherence measure, in order to be doxastically truth conducive, must assign agreement on propositions with antecedent probability 1 maximum degree of coherence. C_2 fails to satisfy this requirement. By contrast, C_1 does satisfy it, taking on its maximum value 1 for any collection of coinciding propositions, tautologies included.

---

[20] See Klein and Warfield, "What Price Coherence?" p. 130; and Merricks, pp. 308-09.

[21] Strictly speaking, we do not have to adopt an objective interpretation of probability if the only thing we want to do is escape the threat of collapse. So long as *I* am interesting in the coherence and truth of *your* beliefs, I can use my subjective probabilities. As just mentioned, however, the use of subjective probabilities would make a connection between coherence and *objective* truth more problematic than if objective probabilities were employed.

What is particularly noticeable about definition 2 is that it introduces a new propositional layer into the picture. We now have, on the one hand, propositions like A_1, A_2, and so on, with the intended interpretation that they are about the world (or some interesting part of it) and, on the other hand, propositions like BelA_1, BelA_2, and so on, which are about what someone believes about the world. Coherence pertains to the first level, that is, to the level of content rather than to the level of belief report. (I shall say more in the next section about the relation between belief reports and their contents.)

As was hinted at already in the introduction, although I am primarily interested in the truth conduciveness of coherence as applied to beliefs, the main issues do not pertain essentially to belief reports, but to reports in general.[23] A statement to the fact that a person believes a given proposition can be taken as a report on that proposition, in the sense of (putatively) indicating its truth. For instance, learning that Albert Einstein believed that A, where A is some proposition about physics, would make most of us more inclined to think that A is true. (If in doubt about this particular example, the reader is encouraged to replace Einstein by an authority of her own choice.) Belief reports are just one type of report. Smith's saying that A or

---

[22] Similar observations are made in Shogenji's "Is Coherence Truth-Conducive?" and my "Why Coherence Is Not Truth-Conducive."

[23] This point is stressed by Lewis: "It may also serve to emphasize the importance of congruence in the confirmation of empirical beliefs if we observe in how large a measure the final basis of credibility must be found in evidence having the character of 'reports' of one kind or another—reports of the senses, reports of memory, reports of other persons—and this label 'report' is appropriate just because such items do not fully authenticate what is 'reported'" (AKV, 357).

Jones's remembering that *B* are examples of other types. In general, we are presented with a number of reports, and we are interested in the relation between the coherence of the report contents and their (conditional) joint probability. The most general question, then, is whether more coherent propositions are more likely jointly to be true among propositions that have been reported individually to be so. It should be mentioned that the sort of interpretative difficulties that arose in connection with belief reports do not arise for reports in general. My subjective probability that *A*, given that I believe that *A* is true, is always 1, a fact which was seen above to lead to trivialization in combination with the conditional account of truth conduciveness. But my subjective probability that *A*, given that I (seem to) remember that *A* is true, need not be 1. For similar reasons, subjective vocabulary can be unproblematically adopted in most other contexts involving nondoxastic reports.

In an attempt to spell out the sense of truth conduciveness underlying the epistemology of Bonjour, Charles B. Cross (*op. cit.*) arrives at a conditional account similar to mine. Following Cross, we let the relation $J(B,c,s,r)$ stand for '*B* is justified to a degree defined by $\langle c,s,r \rangle$'. The relation holds if and only if "*B* is the conjunction of the members of the current belief set of an actual agent whose belief history has length r and consists of belief sets that have remained coherent to degree c and stable to degree s while satisfying the Observation Requirement" (*op. cit.*, p. 189). In order to satisfy Bonjour's observation requirement, a system of beliefs "must contain laws attributing a high degree of reliability to a reasonable variety of cognitively spontaneous beliefs" (*op. cit.*, p. 144). The latter are roughly beliefs that have been acquired noninferentially. Cross goes on to construe Bonjour's truth-conduciveness claim as follows: if $\langle c_2,s_2,r_2 \rangle$ represents a greater degree of justification than $\langle c_1,s_1,r_1 \rangle$, then $\mathrm{P}(B_2|J(B_2,c_2,s_2,r_2)) > P(B_1|J(B_1,c_1,s_2,r_1))$.

Note that Cross conditionalizes, just as we did, on the propositions being believed by an actual agent. But he explicates the truth conduciveness of a certain notion of (degree of) justification seen as applicable to a package of three parameters, including not only the degree of coherence but also two additional parameters: the degree of stability and the length of the belief history. While Cross may well be right in that this conception corresponds to Bonjour's intentions, he has not answered the question whether, and in what sense, coherence *per se* is truth conducive (nor did he intend to). Moreover, it seems to me sound methodology to study the truth conduciveness of the components before studying the truth conduciveness of a whole package. The Bonjour/Cross strategy thus reverses the natural order of inquiry. What I shall say in the next section will raise the suspicion that the observation requirement is actually doing all the work in the resulting theory.

III INDEPENDENCE AND PARTIAL RELIABILITY

It is implausible to think that coherence could be truth conducive, in the doxastic sense, in the absence of further conditions—for example, conditions connecting the belief reports with their contents. I am now concerned with the question of what these additional conditions might be. Before considering some more substantial conditions, I note that the joint probability of the contents of the reports must not be 0 since nothing, coherence being no exception, could affect the joint probability in that case. For the same reason, the joint probability of the contents must not be 1. The satisfaction of these two conditions will be tacitly assumed in the following.

According to Lewis, a necessary condition for coherence to raise credibility is that the reporters tell their story independently (AKV, 346). For if a coherent set is "fabricated out of whole cloth, the way a novelist writes a novel, or if it should be set up as an elaborate hypothesis ad hoc by some theorist whose enthusiasm runs away with his judgment, such congruence would be no evidence of fact" (AKV, 352). If it were such evidence, then "unreliable reporters would be working in the interest of truth if they got together and fudged their stories into agreement" (AKV, 352), a notion which Lewis, quite rightly, dismisses as absurd.

Taking Lewis's remarks on independence as his point of departure, Bonjour writes:

> [a]s long as we are confident that the reports of the various witnesses are genuinely independent of each other, a high enough degree of coherence among them will eventually dictate the hypothesis of truth telling as the only available explanation of their agreement.... And by the same token, so long as cognitively spontaneous beliefs are genuinely independent of each other, their agreement will eventually generate credibility... (*op. cit.*, p. 148).

Bonjour is saying that witnesses have to be independent in order for their agreement to have any effect on the probability of what they agree upon. He is adding that the same holds for cognitively spontaneous beliefs. Coherence among such beliefs has a positive effect on their credibility only to the extent that they are independent of each other. We recall that cognitively spontaneous beliefs are acquired noninferentially, which makes it plausible to think that they should indeed normally be independently held, 'normally' because we cannot rule out altogether the possibility that there be common causal factors behind such beliefs.

We may safely conclude that coherence is not truth conducive if the reports are entirely dependent on each other; in such cases, more coherence does not imply a higher posterior joint probability.[24] On the other hand, it is implausible to require full independence for coherence to have the desirable effect; intuitively, a tiny influence of one report on the other does not cancel out the effect of coherence entirely, although it does make that effect less pronounced. Thus, some degree of dependence is compatible with coherence raising the joint probability, but the raise will be less significant than it would have been, had the reports been less dependent.

Let us turn to the reliability issue. One rarely recognized requirement is that the reports must not be fully reliable. That goes for belief reports in particular. If, for each proposition *A*, the probability of *A*, given that *S* believes that *A*, is 1, then the joint probability of the contents of *S*'s beliefs will also be 1, however coherent or incoherent those contents are. If fully reliable reports are admitted, then coherence is not truth conducive. Coherence is impotent for fully reliable believers, but they, of course, have no use for it anyway.

Lewis requires the reports to be, as he puts it, "relatively unreliable" (AKV, 346) for coherence to have any effect on the credibility of the reported information. His choice of term indicates that he wants to rule out full reliability, among other things. Lewis also intends to disqualify the possibility of a report's being entirely disconnected from its content. According to Lewis: "For any of the reports taken singly, the extent to which it confirms what is reported may be slight" (AKV, 346). But it must not be zero; he writes, in his discussion of memory reports, that "If...there were no initial presumption attaching to the mnemically presented...then no extent of congruity with other such items would give rise to any eventual credibility" (AKV, 357).

I prefer 'partial reliability' to Lewis's expression 'relative unreliability', and will use the former in the following. Thus a partially reliable report lies strictly between full reliability and irrelevance. Coherence, according to this view, is effective only if, for each report *E* with content A, $1 > P(A|E) > P(A)$.[25] Applying this to belief reports, we should require that $1 > P(A|\text{Bel}A) > P(A)$; that is, we should require that the presence of a belief support the content of the belief, though without fully authenticating that content.

---

[24] The case of witnesses copying other witnesses' statements is discussed in Alvin Goldman, "Experts: Which Ones Should You Trust?" *Philosophy and Phenomenological Research*, LXIII, 1 (July 2001): 85-110. He shows, using Bayesian reasoning, that the addition of such witnesses has no effect on the probability of what is being said.

We have seen that Lewis is, once again, giving expression to the common-sense view: coherence of completely useless information does not do a thing for the credibility of that information. Coherence cannot create credibility out of nothing, but can at best amplify credibility that is already there. This notwithstanding, Lewis's position on reliability has been contested, indeed by Bonjour, who writes: "What Lewis does not see, however, is that his own [witness] example shows quite convincingly that no antecedent degree of warrant or credibility is required" (*op. cit.*, p. 148). On one plausible interpretation, Bonjour is saying that coherence is truth conducive also when the reports, taken individually, are irrelevant to what is reported.[26]

In an attempt to resolve the dispute between Lewis and Bonjour, I shall give a formal argument to the effect that, unlike what Bonjour claims, independence and complete unreliability are not sufficient for coherence to be truth conducive.[27] More specifically, I shall argue that full independence should be given a certain exact interpretation and then show that coherence fails to be truth conducive under the assumptions of full independence and complete unreliability. I have argued elsewhere that independence, in this context, should be explicated as conditional independence.[28] What follows is a detailed defense of this proposal.

The idea is that reports are independent just in case they are directly influenced only by the respective facts they report on: they are not directly influenced by what other reporters have to say, nor are they directly influenced by facts other than the ones they report on. It will be helpful to refer to an example throughout this section.

---

[25] Note that $1 > P(A|E) > P(A)$ is equivalent to $P(E|A) > P(E|\neg A) > 0$, provided that $1 > P(A) > 0$.

[26] Alternative interpretations of Bonjour's claim are discussed in Bovens and Olsson.

[27] What follows is a generalization of an argument due to Michael Huemer—see his "Probability and Coherence Justification," *The Southern Journal of Philosophy*, xxxv (1997): 463-72. He proves a similar result for the simpler setting in which both witnesses say the same thing.

[28] See Bovens and Olsson. For more on conditional independence, see Wolfgang Spohn, "Stochastic Independence, Causal Independence, and Shieldability," *Journal of Philosophical Logic*, IX (1980): 73-99. Underlying the theory of Bayesian networks, a currently expansive branch of computer science, conditional independence is already a well-established concept of independence. For a standard text on Bayesian networks, see Judea Pearl, *Probabilistic Reasoning in Intelligent Systems* (San Francisco: Kaufmann, 1988).

(It will be clear, I hope, that nothing hinges on the peculiarities of the example.) Suppose there has been a robbery with Robert's being one of the suspects. There are two witnesses, Helen and Peter, available for questioning. Consider the following propositions:

A_1 = Robert has a gun.

A_2 = Robert is a professional criminal.

E_1 = Helen says that Robert has a gun.

E_2 = Peter says that Robert is a professional criminal.

Applying the foregoing explanation, E_1 and E_2 are independent reports on A_1 and A_2, respectively, just in case E_1 is directly influenced only by A_1, and E_2 is directly influenced only by A_2. This does not mean that there could not be other, indirect influences among the propositions. Conditional independence is compatible with E_1's being influenced also by other propositions than A_1, in which case, however, the influence must pass via A_1 and hence be indirect.

How do we translate this into probability theory? The trick is to block the communication path to E_1 via A_1 and see if there are still any influences on E_1. There are two ways to block that path: assuming A_1 true or assuming A_1 false, that is, conditioning on A_1 or conditioning on $\neg A_1$. Similarly for E_2 and A_2. Conditional independence is satisfied if there are no residual influences. In the interest of avoiding technicalities, I shall confine myself in the following to the two-report case.

Definition 3: E_1 and E_2 are *independent reports* on A_1 and A_2 if and only if the following hold:

$$\begin{array}{ll}
P(E_2|A_1,A_2)=P(E_2|A_1,A_2,E_1) & P(E_2|\neg A_1,A_2)=P(E_2|\neg A_1,A_2,E_1)\\
P(E_2|A_1,A_2)=P(E_2|A_2) & P(E_2|\neg A_1,A_2)=P(E_2|A_2)\\
P(E_1|A_1,A_2)=P(E_1|A_1) & P(E_1|\neg A_1,A_2)=P(E_1|\neg A_1)\\
P(E_2|A_1,\neg A_2)=P(E_2|A_1,\neg A_2,E_1) & P(E_2|\neg A_1,\neg A_2)=P(E_2|\neg A_1,\neg A_2,E_1)\\
P(E_2|A_1,\neg A_2)=P(E_2|\neg A_2) & P(E_2|\neg A_1,\neg A_2)=P(E_2|\neg A_2)\\
P(E_1|A_1,\neg A_2)=P(E_1|A_1) & P(E_1|\neg A_1,\neg A_2)=P(E_1|\neg A_1)
\end{array}$$

Note that independence thus defined is a four-place relation.

According to the second equation on the left, for instance, if we already know that Robert is a criminal, then hearing in addition that he owns a gun does not change our confidence that Peter will report that Robert is a criminal. Our level of confidence that Peter will deliver such a report is determined once the truth value of its content is known; all other factors have been screened off. As a special case, substituting BelA_1 and BelA_2 for E_1 and E_2, respectively, throughout definition 3 gives us an account of when two belief reports are independent with respect to their contents. On this reading, the same equation says that, if we already know that Robert is a criminal, then learning in addition that he owns a gun does not change our confidence that Peter will *believe* that Robert is a criminal. Our level of confidence that Peter will adopt such a belief is determined once the truth value of the content is known.

There are two other concepts of independence that should be sharply distinguished from conditional independence in the sense just defined. First, one could mean by independence that the *contents* of the reports are independent, in the sense of $P(A_1|A_2) = P(A_1)$. In the Robert case, the contents would be independent, in this sense, if knowing Robert to be a professional criminal would neither add to, nor subtract from, our expectation that he will own a gun. Independence in this sense interferes with coherence and results generally in a low coherence value. For instance, while the C_2-measure can take on any real value greater than or equal to 0, the C_2-value of two content independent propositions is always 1. If content independence were the relevant notion of independence for the purposes of coherence and truth, it would be impossible for two independent witnesses to give agreeing testimonies, which is absurd.

Second, there is independence in the sense of $P(E_1|E_2) = P(E_1)$: knowing that Peter has testified that Robert is a criminal does not provide any information whatsoever as to whether Helen will testify that Robert has a gun. Normally, however, hearing one witness testify to a certain effect raises our confidence that the next witness to be questioned will give an agreeing statement. This kind of dependence is quite compatible with the witnesses being independent in the sense relevant for the purposes of coherence and truth, or so I claim. To see this, suppose that E_1 and E_2 are partially reliable reports on A_1 and A_2, respectively. Even if Helen and Peter base their reports directly only on the respective facts and have never communicated, Peter's testifying that Robert is a professional criminal would raise our expectations that Helen will testify that Robert has a gun. Why? Peter's being partially reliable, his testimony would raise our confidence that Robert actually is a criminal which, in turn, would raise our confidence that Robert actually has a gun (because of the coherence of criminality and gun possession). Helen's being also partially reliable, it now becomes more probable than it was before that she will testify to the effect that Robert has a gun. Hence, after we have heard Peter's positive testimony, we would be slightly more inclined than we were before to expect that Helen will testify that Robert has a gun. And yet the testimonies were assumed to be independent in the relevant sense.

I hope to have made it plausible that the relevant sense of (full) independence is conditional independence and proceed now to the second part of my formal argument for requiring partial reliability, which amounts to showing that coherence is not truth conducive under conditional independence and complete unreliability. Fortunately, this is quickly accomplished.

Observation: suppose that the following hold:
(i) E_1 and E_2 are independent reports on A_1 and A_2
(ii) $P(A_1|E_1) = P(A_1)$ and $P(A_2|E_2) = P(A_2)$
(iii) $A_1 \wedge A_2$, $A_1 \wedge \neg A_2$, $\neg A_1 \wedge A_2$ and $\neg A_1 \wedge \neg A_2$ all have nonzero probability.
Then $P(A_1 \wedge A_2|E_1,E_2) = P(A_1 \wedge A_2)$[29]

The third condition serves to rule out certain uninteresting limiting cases. Thus, it follows (essentially) already from conditional independence that reports that are completely unreliable regarding their contents, when taken singly, fail to have any effect on the joint probability of those contents, when combined.

Now, if BelA_i is substituted for E_i in the observation, it follows that doxastic truth conduciveness collapses into propositional truth conduciveness under (essentially) the conditions of conditional independence and complete unreliability. Having established (in section II), that coherence fails to be propositionally truth conducive, we may conclude that coherence fails to be doxastically truth conducive under (essentially) those conditions. Hence, besides independence,

---

[29] *Proof.* By Bayes's Theorem,

$$(1)\ P(A_1 \wedge A_2|E_1,E_2) = \frac{P(E_1 \wedge E_2|A_1,A_2)P(A_1 \wedge A_2)}{P(E_1 \wedge E_2)}$$

By conditional independence, $P(E_1 \wedge E_2|A_1,A_2) = P(E_1|A_1)P(E_2|A_2)$. It follows from (ii) and familiar probabilistic facts that $P(E_i|A_i) = P(E_i)$, $i = 1,2$. Hence,

$$(2)\ P(E_1 \wedge E_2|A_1,A_2) = P(E_1)P(E_2)$$

By (iii) and the theorem of total probability, $P(E_1 \wedge E_2) = P(E_1 \wedge E_2|A_1,A_2)P(A_1 \wedge A_2) + P(E_1 \wedge E_2|A_1,\neg A_2)P(A_1 \wedge \neg A_2) + P(E_1 \wedge E_2|\neg A_1,A_2)$ $P(\neg A_1 \wedge A_2) + P(E_1 \wedge E_2|\neg A_1,\neg A_2)P(\neg A_1 \wedge \neg A_2)$. By conditional independence, the right-hand side of that equation equals $P(E_1|A_1)P(E_2|A_2)P(A_1 \wedge A_2) + P(E_1|A_1)P(E_2|\neg A_2)P(A_1 \wedge \neg A_2) + P(E_1|\neg A_1)P(E_2|A_2)P(\neg A_1 \wedge A_2) + P(E_1|\neg A_1)$ $P(E_2|\neg A_2)P(\neg A_1 \wedge \neg A_2)$. As already noticed, it follows from (ii) that $P(E_1|A_1) = P(E_1)$ and $P(E_2|A_2) = P(E_2)$. It also follows from (ii) that $P(E_2|\neg A_2) = P(E_2)$ and $P(E_1|\neg A_1) = P(E_1)$. Combining all this yields

$$(3)\ P(E_1 \wedge E_2) = P(E_1)P(E_2)P(A_1 \wedge A_2) + P(E_1)P(E_2)P(A_1 \wedge \neg A_2) + P(E_1)P(E_2)P(\neg A_1 \wedge A_2) + P(E_1)P(E_2)P(\neg A_1 \wedge \neg A_2) = P(E_1)P(E_2).$$

It follows from (1), (2) and (3) that $P(A_1 \wedge A_2|E_1,E_2) = P(A_1 \wedge A_2)$, which ends the proof.

partial reliability is also needed to provide the context within which coherence can do useful, amplifying work.

Lewis suggested an explication of independence strikingly similar to, but not identical with, the one proposed here. Lewis introduces his concept of independence when discussing the corroboration of a hypothesis through its testable consequences. Since the testable consequences of a hypothesis are (putative) indicators of its truth and hence are reports, in the abstract sense, on that hypothesis, his remarks on independence are easily generalized beyond the hypothesis-consequences setting.[30] He considers two propositions, *A* and *B*, to be independent consequences of a hypothesis *H* just in case *A* and *B* are "so related that supposing *H* false, the finding of one of them true would not increase the probability of the other" (AKV, 344). Lewis restates this idea several times, with little variation, writing for instance that "in general, the consequences of a hypothesis are independent only in the sense that the establishment of one does not increase the probability of another on the assumption that the hypothesis is *false*" (AKV, 349, footnote 6). It was hence clear to Lewis that the relevant notion of independence is of a conditional nature. But he insisted on taking into account only independence statements conditional on the falsity of the hypothesis, leaving out independence statements conditional on its truth. While he was on the right track, the kind of independence he arrived at is ultimately too weak for the purposes of coherence and truth.

IV. TRUTH CONDUCIVENESS CETERIS PARIBUS

I shall now discuss yet another condition that must be satisfied for coherence to stand a reasonable chance of being truth conducive. Suppose that we want to evaluate the truth conduciveness of a given measure of coherence. We now find that the propositions reported by one set of witnesses are more coherent, according to the measure, than the propositions reported by another set of witnesses, and also that the conditions we have discussed so far are satisfied; that is, in each scenario the witnesses are partially reliable and independent. But because the witnesses delivering the more coherent set are less reliable, or more dependent, than the witnesses delivering the less coherent set, the posterior joint probability turns out to be higher for the latter. Should we take this as a counterexample to the truth conduciveness of the coherence measure in question? Presumably, we should not. Supposedly, the claim that more coherence implies a higher likelihood of truth is a ceteris paribus claim: more coherence implies a higher likelihood of truth *if everything else is equal*. In particular, neither the degree of reliability nor the degree of independence should be allowed to vary across scenarios under comparison.

---

[30] Lewis explains the relevant notion of consequence as follows: "when a hypothesis *H* is said to have consequences *C*, what typically is meant is that *H*, together with other statements which may reasonably be assumed, gives a high probability of *C*" (AKV, 343, notation adapted). He adds, in footnote 5 on the same page, that "[i]t is not even essential that such a probability be high," that is to say, "in general the same principle will apply wherever *C* is more probable than not."

Shogenji,[31] too, notes that certain "confounding factors" have to be kept fixed in the evaluation of the truth conduciveness of coherence. He does not mention reliability and independence in this connection, which is not surprising, considering the fact that he (as we have seen) is concerned with the propositional conception of truth conduciveness and not with the doxastic or testimonial conception. In the propositional one-layer model, issues of reliability and independence (in the relevant sense) do not arise. Instead, Shogenji claims that total individual strength is a confounding factor with respect to coherence. In the evaluation of the truth conduciveness of the latter, the former has to be held fixed.

Before evaluating Shogenji's contention about strength, let us take a look at ceteris paribus conditions from an abstract point of view. As Robert L. Frazier[32] notes, other things are equal when there is only one candidate for being an influence on some outcome. By requiring other influences to be excluded, we are trying to isolate the target property. The truth of a statement with a ceteris paribus clause depends on what happens when this isolation is achieved. Frazier (*ibid.*, p. 119) observes that the properties that have been thus singled out must be capable of independent variation; otherwise, they are not really separable. He goes on to describe what it would mean, more precisely, for two properties P_1 and P_2 to be *in*capable of independent variation. That would mean, he writes, that whenever an object has P_1 to a particular degree, say, n, the object also has P_2 to a particular degree, say, m, so that "when one occurs to a particular degree this is then a certain sign that the other occurs to a particular degree" (*ibid.*).

While I agree with Frazier that the requirement of independent variation would be violated in this case, I do not think that this sort of situation represents the only type in which it would be violated. Rather, this would be an extreme case. For suppose that whenever an object has P_1 to a particular degree, say, n, it has P_2 either to a degree m *or* to a degree m'. Intuitively, this would also amount to lack of independent variation, provided, of course, that P_2 could take on more than just the two values m and m' before the value of P_1 was fixed. In general, the requirement of independent variation is violated, I take it, if fixing the value of the one property imposes limitations on the extent to which the other property can consistently vary.[33]

---

[31] See his "Is Coherence Truth-Conducive?"

[32] "Moral Relevance and *Ceteris Paribus* Principles," *Ratio*, VII, 2 (September 1995): 113-27, especially p. 114

In the present case, we are trying to isolate one property—namely, coherence—from other properties that may have an influence on the joint probability of a set of beliefs or testimonies. It can now be seen that, if we use Shogenji's own measure of coherence (C_2), we cannot include strength among the properties that should be held fixed. The reason is that strength and coherence are not independent properties in the relevant sense. The reader might have suspected this already on the grounds that the strength figures as the denominator in the definition of that measure. To prove the case, note that, if the strength, that is, $P(A) \times P(B)$, is not fixed, $C_2(A,B)$ has no upper bound; it can take on any real value greater than, or equal to, 0. If we keep the strength fixed, however, $C_2(A,B)$ becomes bounded. For instance, if the strength is assigned a value of 9:10, then the upper bound is 10:9. Thus, keeping the strength fixed imposes limitations on the extent to which the degree of coherence can consistently vary, in violation of the requirement of independent variation.[34]

---

[33] Independent variation, in this sense, is compatible with the instantiation of one property to a particular degree making the instantiation of another property to a particular degree highly likely, so long as the resulting likelihood is less than 1.

[34] In Bovens and Olsson (*op. cit.*, p. 691) we argued that the antecedent joint probability should be held fixed in assessing whether coherence is truth conducive, referring to the following example. Let us assume that two independent witnesses testify to the effect that a dog died due to witchcraft, what was antecedently very improbable. Suppose also that two other independent witnesses provide us with reports that are not in complete agreement but still not mutually exclusive—for example, that the dog died due to cancer and due to pneumonia, respectively. All reporters are initially considered partially reliable to the same degree. Clearly, the former reports are more coherent than the latter reports. And yet the posterior (joint) probability that the dog died due to witchcraft is still lower than that it died due to the combined effect of cancer and pneumonia. Why? Because of the difference in priors. This example was taken to illustrate the thesis that the prior joint probability may influence the posterior joint probability and that we need to keep it fixed on the grounds that we need to keep fixed all factors, other than coherence, that may affect the posterior joint probability But the question is whether the prior joint probability is clearly separable from coherence, so that we may legitimately speak of them as two distinct factors. For reasons similar to those advanced here against fixing the strength, I no longer think that prior joint probability can be isolated from coherence. Fixing the prior may lead to violations of the requirement of independent variation. The effect of varying the prior

By contrast, holding the degree of reliability or independence fixed is in perfect compliance with the requirement of independent variation. Fixing these aspects of scenarios does not mean limiting in any way what degree of coherence can be consistently attributed. For instance, if we learn that the each witness tells the truth sixty percent of the time, this information does not put us in a better position than we were before to eliminate coherence values.[35]

It is not surprising that coherence and reliability should admit of independent variation: while the former is a property on the content level, the latter concerns the relation between contents and reports. Hence, the former is, in a sense, "horizontal" and the latter "vertical." On the other hand, coherence and total individual strength are both properties on the horizontal content level, which is why we should expect them to interfere, as indeed they do.

Should we hold fixed the size of an information set in the evaluation of the truth conduciveness of coherence? That would be reasonable if size were something that could be isolated from coherence. To be sure, just by focusing on information sets of a certain size, one has not thereby in any way restricted the assignment of a degree of coherence. On the other hand, when the size is changed as the result of the addition of one or more items of information (for example, further testimonies), this usually also affects the degree of coherence. Size is, in this sense, not completely separable from coherence. By contrast, changing the reliability or independence of the witnesses never brings with it a change in coherence. It seems to me that the ceteris paribus conception requires this strong sense of separation illustrated by reliability and independence. If this is true, size should not be held fixed.

If size should not be held fixed, then C_1 is not truth conducive ceteris paribus. Suppose that we have bet money on the outcome of casting a die but are unable to be present to confirm the result. We decide to consult informants who were present. Consider two scenarios: in the first scenario, two persons were present to watch the event,

---

probability of what is being said on its posterior probability is studied in my "Corroborating Testimony, Probability and Surprise" (to appear in *The British Journal for the Philosophy of Science*) in the simpler context involving two witnesses testifying to the same effect.

[35] If we were to learn that the witnesses *always* tell the truth, that would indeed tell us something conclusive about the coherence: it would enable us to exclude the case of their delivering incompatible testimonies. Hence, it would seem that reliability is not independent of coherence after all. Recall, however, that in requiring the witnesses to be partially reliable, we have already ruled out perfectly reliable witnesses from consideration.

each reporting that 'One' came up. In the second scenario, 100 persons were present, out of which 99 report that 'One' came up, and the one remaining that either 'One' or 'Two' did. In the first scenario, there is perfect agreement among the witness statements (and hence a maximum C_1-value), in the second not (and hence a less than maximum C_1-value). Yet it is clear that we would, everything else equal, be more confident that 'One' came up in the second, less coherent scenario. Hence, a lower degree of C_1-coherence may be counterbalanced by a larger number of testimonies. This casts doubts on the doxastic case as well, as it raises the suspicion that a higher degree of C_1-coherence could be outweighed by a sufficient increase in the number of beliefs.

This negative result depends on the particular coherence measure. If instead C_2 is used, for example, we cannot conclude that the degree of coherence (agreement) is greater in the first scenario (with two reports) than in the second (with 100 reports). In fact, the second scenario will exhibit a much higher degree of C_2-coherence than the first. To see this, recall the quotient in the definition of C_2. While the numerator of that quotient (that is, the joint probability) will be the same for the two scenarios, the denominator (that is, the total individual strength) will be much smaller for the second scenario, thus making the value of the whole expression much higher.

The problem of coherence and truth, or at least one clear and manageable subproblem, is finding a (nontrivial) measure of coherence that is doxastically truth conducive, ceteris paribus under (essentially) the conditions of partial reliability and independence. At the same time, it must be added that there are reasons for pessimism concerning the existence of such measures. The measures we started out with seemed to capture important intuitions about coherence reasonably well but they turned out, nonetheless, not to be truth conducive in this sense. Neither C_1 nor C_2 is truth conducive ceteris paribus under the conditions in question. C_1, as we just observed, has a problem of size. We have noticed before (in section II), that C_2 has difficulties with strength (specificity).

V. ON THE POSSIBILITY OF READING OFF THE POSTERIOR FROM THE COHERENCE

The kind of relation between coherence and truth which has been the subject of our discussion so far is of a comparative nature. Given that set S is more coherent than set S', can we conclude that S is also more likely than S' to be true as a whole? We found that this question could hardly be answered in the affirmative unless it is qualified in certain substantial respects. I shall now try to make plausible that there is nothing like a degree of coherence that would allow us to read off the likelihood of truth of a given set from its degree of coherence. While it remains an open question whether a *higher* degree of coherence implies a *higher* likelihood of truth (given the appropriate qualifications), we can be confident that a *high* degree of coherence does not necessarily imply a *high* likelihood of truth. Once again, my point of departure will be Lewis's view.

In his attempt to provide a validation of memory, Lewis argues that we can know, on a priori grounds, that the partial reliability requirement is satisfied in that context.[36] But this is also all we can say. For all we know, the initial credibility (reliability) of our memory may be high, low, or anything in between. Lewis expresses this contention by saying that "the degree of this initial credibility...is hardly assignable" (AKV, 357). According to Lewis, then, our initial condition as regards the reliability of our memory is one of partial ignorance.

Lewis goes on to advance two theses about the relation between coherence and likelihood of truth. Unfortunately, what says in this connection is perplexing, as he appears to contradict himself.

Lewis claims that it does not matter that the degree of initial credibility is not assignable because—and this one of his two theses—"it does not need to be assigned" (AKV, 357). Now, whether something does or does not need to be done is relative to some purpose. This raises the question of what purpose Lewis could have had in mind here. The answer seems clear enough: for the purpose of determining the posterior (eventual) probability of a given set of reported propositions, the initial credibility does not need to be assigned. It is sufficient, for that purpose, to know the degree of coherence pertaining to that set. Thus, the claim is that we can read off the posterior from the degree of coherence, without having to take the initial credibility of the information into account, so long as we know that this credibility is positive.

---

[36] According to Lewis, mnemic representation affords some probability of past fact already before any examination as to coherence. Such credibility "is constitutive of our sense of the only reality by reference to which empirical judgments could have either truth or falsity or any meaning at all" (AKV, 361). Lewis's point is that without this assumption our very notion of empirical fact would not make sense, and so there could be no (empirical) basis from which this assumption could be questioned. The assumption in question "is an analytic statement which can only be repudiated on premises which already imply it" (AKV, 361), and skepticism about its validity is ultimately self-defeating.

This thesis is also stated as follows:

> Take the case of the unreliable observers who agree about what they report. In spite of the antecedent improbability of any item of such report, when taken separately, it may become practically certain, in a favorable case, *merely* through congruent relations to other such items, which would be similarly improbable when separately considered (AKV, 352; my emphasis).

Here is another passage in the same spirit:

> *All that is needed* is the initial assumption that the mere fact of present rememberings renders what is thus memorially present in some degree credible. For the rest, the congruence of such items with one another and with present sense experience will be capable of establishing an eventual high credibility, often approximating to certainty, for those items which stand together in extensive relations of such congruence (AKV, 354; my emphasis).

Lewis is here clearly committed to the thesis that, given the initial credibility of our memories, a high enough degree of coherence ("congruence") will *by itself* establish a high posterior probability ("eventual high credibility").

And yet Lewis also writes, in effect, that the posterior probability does *not* depend only on the degree of coherence:

> A larger or a smaller such initial probability would have no appreciable effect upon the eventually determinable probabilities in question *beyond that of a difference in the extent of congruity with other mnemic items and with sense presentation which would be required for building up eventual probabilities sufficient for rational and practical reliance* (AKV, 357; my emphasis).

What Lewis is here mentioning in passing, as if it were of no particular significance to the issue at hand, is actually the crucial thing. For he is saying that the posterior probability is a function not only of the degree of coherence ("extent of congruity") but also of the level of initial credibility. More precisely, the lower the initial credibility is, the higher the degree of coherence required for building up posterior probabilities sufficient for "rational and practical reliance." If our memories are initially already rather reliable, they do not have to agree very much in order for us to be willing to rely on them. On the other hand, if they are initially highly dubious, striking agreement would be needed to achieve the same effect. All this presupposes, of course, that the independence condition is satisfied.

Lewis's second thesis is clearly the correct one. In fact, we took its validity for granted in our discussion of ceteris paribus clauses when we argued (in section IV) that the degree of reliability should be held fixed in the evaluation of the truth conduciveness of coherence. This has grave consequences for Lewis's epistemological enterprise. Since the posterior does not depend only on the degree of coherence but also on the initial credibility, in order to know whether we can, rationally and practically, rely on the information at hand, we have to know what the initial credibility is more precisely; it does not suffice just to know that it is positive. Otherwise, we could not be confident that the degree of coherence pertaining to the given information set is high enough. Hence, for the purposes of determining the posterior probability, the initial credibility does need to be assigned. But, says Lewis, it cannot be assigned. It follows that, unless the antecedent joint probability was already 1, we can never be confident that our probabilities are sufficient for rational and practical reliance. The net result is a form of skepticism that would have been unacceptable to Lewis.[37]

Apart from their ramifications for Lewis's theory, these considerations strongly suggest, if not prove, that there is no measure, probabilistic or not, definable solely in terms of the contents of beliefs, memories, or testimonies which would allow us to conclude anything interesting about how high the posterior probability is in absolute terms. In order to draw such conclusions, we also need to know (at least) the initial reliability of the reports. In the absence of such information, we may at best draw the comparative conclusion that more coherence implies a higher posterior probability.

We have already seen (in section III), that Bonjour's coherence theory is based on the false probabilistic assumption that, if individually useless data are independent, they can become collectively useful, in the sense of yielding, when combined, a high posterior probability. Lewis wisely rejected this notion, pointing out that coherence has an effect only if the individual data have some degree of initial credibility. But he, too, was unrealistically optimistic in his assessment of what coherence can actually accomplish. His (weak) foundationalism—his attempt to provide a "final validation of empirical knowledge by reference to ultimate data in some sense presently given, and to the congruence of such data" (AKV, 356), given our supposed partial ignorance regarding the initial credibility of those data—shares the deficiency of Bonjour's theory of being founded on an incorrect probabilistic supposition. In Lewis's case, the erroneous, if essential, assumption is that posterior probability is determined by coherence alone, once the minimum conditions of partial reliability and independence are satisfied.[38]

---

[37] Lewis's attempted justification of memory beliefs is studied in greater detail in a forthcoming paper by me and Shogenji entitled "C. I. Lewis on Congruence and Probability."

Summing up, we have seen that theories that rely on the truth conduciveness of coherence are frequently misinterpreted as advancing a propositional conception of the truth connection. Instead, truth conduciveness should be understood in a doxastic, or testimonial, sense. The general conclusion, once this initial confusion has been sorted out, is that there is nevertheless at best a very weak connection between coherence and (likelihood of) truth. In particular, we should not hope to establish anything more substantial than the comparative statement that more coherence implies a higher likelihood of truth. Moreover, the comparative statement, in order to stand a reasonable chance of being valid, should be further qualified as involving a ceteris paribus clause and as presupposing the obtaining of certain favorable circumstances—mainly partial reliability and independence. Whether we can establish even this much remains an open question. We also saw that Bonjour's coherence theory and Lewis's weak foundationalism turn out to be untenable already on the grounds that they rely, each in its own essential way, on probabilistic assumptions that are simply incorrect.

ERIK J. OLSSON

University of Constance

---

[38] According to Huemer, "Lewis' fairly modest claim seems to have been just this: that you can get a high degree of confirmation as a result of the agreement of the witnesses even when each witness has only a low degree of credibility considered by himself" (*op. cit.*, p. 468). Huemer takes the truth of the "modest claim" to warrant the conclusion that "weak foundationalism should be taken seriously as an epistemological position" (p. 471). Lewis, indeed, claimed that it is possible to obtain a high degree of confirmation in this way, if the circumstances are favorable. Unlike what Huemer seems to think, this fairly uncontroversial statement (as I hope to have made clear above), does not exhaust Lewis's weak foundationalism.

NEW BOOKS: ANTHOLOGIES

ALLEN, RICHARD, and MALCOLM TURVEY, eds.: *Wittgenstein, Theory, and the Arts.* New York: Routledge, 2001. x, 302 p. Cloth $75.00.

ANTONY, LOUISE M., and CHARLOTTE E. WITT, eds.: *A Mind of One's Own: Feminist Essays on Reason and Objectivity, Second Edition.* Boulder, CO: Westview, 2002. xix, 443 p. Cloth $90.00, paper $32.00.

ARRINGTON, ROBERT L., and MARK ADDIS, eds.: *Wittgenstein and Philosophy of Religion.* New York: Routledge, 2001. xvi, 187 p. Cloth $85.00.

ASCHHEIM, STEVEN E., ed.: *Hannah Arendt in Jerusalem.* Berkeley: California UP, 2001. xii, 428 p. Paper $22.50.

BAKER, KEITH MICHAEL, and PETER HANNS REILL, eds.: *What's Left of Enlightenment? A Postmodern Question.* Stanford: University Press, 2001. ix, 203 p. Cloth $45.00, paper $19.95.

BALASHOV, YURI, and ALEX ROSENBERG, eds.: *Philosophy of Science: Contemporary Readings.* New York: Routledge, 2002. xiii, 522 p. Cloth $65.00, paper $19.95.

BECHTEL, WILLIAM, PETE MANDIK, JENNIFER MUNDALE, and ROBERT S. STUFFLEBEAM, eds.: *Philosophy and the Neurosciences: A Reader.* Malden, MA: Blackwell, 2001. xiii, 482 p. Cloth $74.95, paper $39.95.

BECKER, LAWRENCE C., and CHARLOTTE B. BECKER, eds.: *Encyclopedia of Ethics, Second Edition, Volumes I-III.* New York: Routledge, 2001. xxxv, 1977 p. Cloth $295.00.

BEINER, RONALD, and JENNIFER NEDELSKY, eds.: *Judgment, Imagination, and Politics: Themes from Kant and Arendt.* Lanham, MD: Rowman and Littlefield, 2001. xxvi, 319 p. Cloth $70.00, paper $26.95.

BOWIE, NORMAN E., ed.: *The Blackwell Guide to Business Ethics.* Malden, MA: Blackwell, 2002. x, 363 p. Cloth $69.95.

BRADY, EMILY, and JERROLD LEVINSON, eds.: *Aesthetic Concepts: Essays after Sibley.* New York: Oxford, 2001. ix, 239 p. Cloth $49.95.

BRANQUINHO, JOÃO, ed.: *The Foundations of Cognitive Science.* New York: Oxford, 2001. xlvii, 235 p. Paper $19.95.

BULHOF, ILSE N., and LAUREN TEN KATE, eds.: *Flight of the Gods: Philosophical Perspectives on Negative Theology.* New York: Fordham, 2000. xiii, 444 p. Cloth $39.95, paper $19.95.

BUNGE, MARIO: *Scientific Realism.* MARTIN MAHNER, ed. Amherst, NY: Prometheus, 2001. 438 p. Cloth $32.00

BURLEY, JUSTINE, and JOHN HARRIS, eds.: *A Companion to Genethics.* Malden, MA: Blackwell, 2002. xiv, 489 p. Cloth $99.95.

CAHN, STEVEN M., ed.: *Classics of Political and Moral Philosophy.* New York: Oxford, 2002. xiii, 1211 p. Paper $35.00.

CAPUTO, JOHN D., MARK DOOLEY, and MICHAEL J. SCANLON, eds.: *Questioning God.* Bloomington: Indiana UP, 2001. viii, 379 p. Cloth $59.95, paper $24.95.

CHAMBERS, SIMONE, and WILL KYMLICKA, eds.: *Alternative Conceptions of Civil Society*. Princeton: University Press, 2002. vi, 237 p. Cloth $59.50, paper $17.95.

CHANTER, TINA, ed.: *Feminist Interpretations of Emmanuel Levinas*. University Park: Penn State UP, 2001. xv, 272 p. Cloth $65.00, paper $25.00.

CORCORAN, KEVIN, ed.: *Soul, Body, and Survival: Essays on the Metaphysics of Human Persons*. Ithaca: Cornell, 2001. vii, 252 p. Cloth $45.00, paper $19.95.

DAVENPORT, JOHN J., and ANTHONY RUDD, eds.: *Kierkegaard after MacIntyre: Essays on Freedom, Narrative, and Virtue*. Chicago: Open Court, 2001. xxxix, 363 p. Cloth $56.95, paper $26.95.

DAVIS, MICHAEL, and ANDREW STARK, eds.: *Conflict of Interest in the Professions*. New York: Oxford, 2001. vii, 355 p. Cloth $55.00.

DOYLE, LAURA, ed.: *Bodies of Resistance: New Phenomenologies of Politics, Agency, and Culture*. Evanston: Northwestern, 2001. xxxiv, 305 p. Cloth $89.95, paper $29.95.

DYZENHAUS, DAVID, and ARTHUR RIPSTEIN, eds.: *Law and Morality: Readings in Legal Philosophy, Second Edition*. Toronto: University Press, 2001. xv, 1061 p. Cloth $100.00, paper $45.00.

EGONSSON, DAN, JONAS JOSEFSSON, BJÖRN PETERSSON, and TONI RØNNOW-RASMUSSEN, eds.: *Exploring Practical Philosophy: From Action to Values*. Burlington, VT: Ashgate, 2001. vii, 197 p. Cloth $79.95.

ELIO, RENÉE, ed: *Common Sense, Reasoning, and Rationality*. New York: Oxford, 2002. x, 278 p. Paper $29.95.

ELLIOTT, CARL, ed.: *Slow Cures and Bad Philosophers: Essays on Wittgenstein, Medicine, and Bioethics*. Durham, NC: Duke, 2001. x, 197 p. Cloth $54.95, paper $18.95.

EMMANUEL, STEVEN M., and PATRICK GOOLD, eds.: *Modern Philosophy: From Descartes to Nietzsche: An Anthology*. Malden, MA: Blackwell, 2002. xvii, 525 p. Cloth $74.95.

ESSED, PHILOMENA, and DAVID THEO GOLDBERG, eds.: *Race Critical Theories: Text and Context*. Malden, MA: Blackwell, 2001. xix, 537 p. Cloth $64.95, paper $29.95.

ESTLUND, DAVID, ed.: *Democracy*. Malden, MA: Blackwell, 2002. viii, 342 p. Paper $24.95.

FAIRWEATHER, ABROL, and LINDA ZAGZEBSKI, eds.: *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility*. New York: Oxford, 2001. ix, 251 p. Cloth $45.00.

FESTENSTEIN, MATTHEW, and SIMON THOMPSON, eds.: *Richard Rorty: Critical Dialogues*. Malden, MA: Blackwell, 2001. ix, 241 p. Cloth $59.95, paper $22.95.

FRIESE, HEIDRUN, ed.: *The Moment: Time and Rupture in Modern Thought*. Liverpool: University Press, 2001. 209 p. Cloth $67.95.

GALAVOTTI, MARIA CARLA, PATRICK SUPPES, and DOMENICO COSTANTINI, eds.: *Stochastic Causality*. Stanford: CSLI, 2001. xii, 275 p. Cloth $65.00, paper $25.00.

GANSSLE, GREGORY E., and DAVID M. WOODRUFF, eds.: *God and Time: Essays on the Divine Nature.* New York: Oxford, 2002. xi, 252 p. Cloth $49.95.

GASKIN, RICHARD, ed.: *Grammar in Early Twentieth-Century Philosophy.* New York: Routledge, 2001. x, 258 p. Cloth $115.00.

GIBSON, NIGEL, and ANDREW RUBIN, eds.: *Adorno: A Critical Reader.* Malden, MA: Blackwell, 2002. xiii, 442 p. Paper $29.95.

GILLETT, CARL, and BARRY LOEWER, eds.: *Physicalism and Its Discontents.* New York: Cambridge, 2001. x, 369 p. Cloth $60.00.

GLENDINNING, SIMON, ed.: *Arguing with Derrida.* Malden, MA: Blackwell, 2001. 143 p. Paper $22.95.

GLOCK, HANS-JOHANN, ed.: *Wittgenstein: A Critical Reader.* Malden, MA: Blackwell, 2001. xxvii, 381 p. Cloth $64.95, paper $29.95.

GOLDIE, PETER, ed.: *Understanding Emotions: Mind and Morals.* Burlington, VT: Ashgate, 2002. vii, 135 p. Cloth $69.95, paper $29.95.

HOERL, CHRISTOPH, and TERESA MCCORMACK, eds.: *Time and Memory: Issues in Philosophy and Psychology.* New York: Oxford, 2001. xiii, 419 p. Paper $26.00.

HOHENDAHL, PETER UWE, and JAIMEY FISHER, eds.: *Critical Theory: Current State and Future Prospects.* New York: Berghahn, 2001. x, 317 p. Cloth $59.95, paper $25.00.

HOLLAND, NANCY J., and PATRICIA HUNTINGTON, eds.: *Feminist Interpretations of Martin Heidegger.* University Park: Penn State UP, 2001. xv, 399 p. Cloth $75.00, paper $28.50.

HOLZBERGER, WILLIAM G., ed.: *The Letters of George Santayana: Book One, [1868]-1909.* Cambridge: MIT, 2001. lxiv, 582 p. Cloth $58.00; *The Letters of George Santayana: Book Two, 1910-1920.* Cambridge: MIT, 2002. lxxiii, 576 p. Cloth $55.00.

INGRAM, DAVID, ed.: *The Political.* Malden, MA: Blackwell, 2002. xi, 303 p. Paper $27.95.

JACQUETTE, DALE, ed.: *A Companion to Philosophical Logic.* Malden, MA: Blackwell, 2002. xiii, 816 p. Cloth $124.95; *Philosophy of Logic: An Anthology.* Malden, MA: Blackwell, 2002. xi, 372 p. Cloth $74.95; *Philosophy of Mathematics: An Anthology.* Malden, MA: Blackwell, 2002. xii, 428 p. Cloth $74.95.

KANE, ROBERT, ed.: *Free Will.* Malden, MA: Blackwell, 2002. xii, 310 p. Paper $24.95; *The Oxford Handbook of Free Will.* New York: Oxford, 2002. xvii, 638 p. Cloth $75.00.

KLAGGE, JAMES C., ed.: *Wittgenstein: Biography and Philosophy.* New York: Cambridge, 2001. xv, 272 p. Cloth $54.95, paper $19.95.

KLEMKE, E.D., with HEIMIR GEIRSSON, eds.: *Contemporary Analytic and Linguistic Philosophies, Second Edition.* Amherst, NY: Prometheus, 2000. 495 p. Paper $26.00.

KOISTINEN, OLLI, and JOHN BIRO, eds.: *Spinoza: Metaphysical Themes.* New York: Oxford, 2002. x, 255 p. Cloth $49.95.

KORNBLITH, HILARY, ed.: *Epistemology: Internalism and Externalism.* Malden, MA: Blackwell, 2001. viii, 271 p. Cloth $59.95, paper $24.95.

KOURANY, JANET A., ed.: *The Gender of Science.* Upper Saddle River, NJ: Prentice Hall, 2001. 319 p. Paper $34.80.

KREMER, ELMAR J., and MICHAEL J. LATZER, eds.: *The Problem of Evil in Early Modern Philosophy.* Toronto: University Press, 2001. 179 p. Cloth $60.00.

LAFOLLETTE, HUGH, ed.: *Ethics in Practice: An Anthology, Second Edition.* Malden, MA: Blackwell, 2002. xiv, 680 p. Cloth $74.95.

LARA, MARÍA PÍA, ed.: *Rethinking Evil: Contemporary Perspectives.* Berkeley: California UP, 2001. viii, 307 p. Paper $22.50.

LEITCH, VINCENT B., et al., eds.: *The Norton Anthology of Theory and Criticism.* New York: Norton, 2001. xxxviii, 2624 p. Cloth $67.45.

LUDLOW, PETER, ed.: *Crypto Anarchy, Cyberstates, and Pirate Utopias.* Cambridge: MIT, 2001. xxii, 485 p. Paper $24.95.

MACHAMER, PETER, and MICHAEL SILBERSTEIN, eds.: *The Blackwell Guide to the Philosophy of Science.* Malden, MA: Blackwell, 2002. xi, 347 p. Paper $34.95.

MALPAS, JEFF, ULRICH ARNSWALD, and JENS KERTSCHER, eds.: *Gadamer's Century: Essays in Honor of Hans-Georg Gadamer.* Cambridge: MIT, 2002. xiv, 363 p. Paper $25.00.

MARCUSE, HERBERT: *Towards a Critical Theory of Society.* DOUGLAS KELLNER, ed. New York: Routledge, 2001. x, 242 p. Cloth $35.00.

MARGOLIS, JOSEPH, and TOM ROCKMORE, eds.: *The Philosophy of Interpretation.* Malden, MA: Blackwell, 2000. vi, 236 p. Paper $24.95.

MARRA, MICHAEL F., ed.: *Japanese Hermeneutics: Current Debates on Aesthetics and Interpretation.* Honolulu: Hawaii UP, 2002. xi, 247 p. Cloth $38.00.

MARTINICH, A. P., and DAVID SOSA, eds.: *A Companion to Analytic Philosophy.* Malden, MA: Blackwell, 2001. x, 497 p. Cloth $124.95; *Analytic Philosophy: An Anthology.* Malden, MA: Blackwell, 2001. ix, 517 p. Cloth $74.95.

MCCARTHY, TIMOTHY G., and SEAN C. STIDD, eds.: *Wittgenstein in America.* New York: Oxford, 2001. xi, 280 p. Cloth $45.00.

MILLER, DAVID, and SOHAIL H. HASHMI, eds.: *Boundaries and Justice: Diverse Ethical Perspectives.* Princeton: University Press, 2001. xi, 367 p. Cloth $55.00, paper $19.95.

MILLGRAM, ELIJAH, ed.: *Varieties of Practical Reasoning.* Cambridge: MIT, 2001. xi, 479 p. Paper $35.00.

MOHANTY, J. N.: *Explorations in Philosophy: Essays by J.N. Mohanty, Volume I: Indian Philosophy.* BINA GUPTA, ed. New York: Oxford, 2001. xxxvii, 229 p. Cloth $29.95.

MORRIS, CHRISTOPHER W., and ARTHUR RIPSTEIN, eds.: *Practical Rationality and Preference: Essays for David Gauthier.* New York: Cambridge, 2001. 239 p. Cloth $54.95.

MOU, BO, ed.: *Two Roads to Wisdom? Chinese and Analytic Philosophical Traditions.* Chicago: Open Court, 2001. xvii, 381 p. Paper $34.95.

MUNZER, STEPHEN R., ed.: *New Essays in the Legal and Political Theory of Property.* New York: Cambridge, 2001. xi, 211 p. Cloth $59.95.

PALMER, RICHARD E., ed.: *Gadmer in Conversation.* New Haven: Yale, 2001. vii, 174 p. Cloth $22.50.

PENNOCK, ROBERT T., ed.: *Intelligent Design Creationism and Its Critics: Philosophical, Theological, and Scientific Perspectives.* Cambridge: MIT, 2001. xx, 805 p. Paper $45.00.

PERKINS, ROBERT L., ed.: *International Kierkegaard Commentary: The Concept of Irony, Volume* 2. Macon, GA: Mercer, 2001. x, 427 p. Cloth $35.00.

PLANINC, ZDRAVKO, ed.: *Politics, Philosophy, Writing. Plato's Art of Caring for Souls.* Columbia: Missouri UP, 2001. 261 p. Paper $19.95.

PORTMANN, JOHN, ed.: *In Defense of Sin.* New York: Palgrave, 2001. 308 p. Cloth $23.95.

POSTEMA, GERALD J., ed.: *Philosophy and the Law of Torts.* New York: Cambridge, 2001. ix, 336 p. Cloth $60.00.

PRIMORATZ, IGOR, ed.: *Patriotism.* Amherst, NY: Humanity, 2002. 298 p. Paper $29.00.

RAGLAND, C.P., and SARAH HEIDT, eds.: *What Is Philosophy?* New Haven: Yale, 2001. vii, 196 p. Cloth $30.00, paper $15.00.

RECK, ERICH H., ed.: *From Frege to Wittgenstein: Perspectives on Early Analytic Philosophy.* New York: Oxford, 2002. xv, 470 p. Cloth $65.00.

REGH, WILLIAM, and JAMES BOHMAN, eds.: *Pluralism and the Pragmatic Turn: The Transformation of Critical Theory: Essays in Honor of Thomas McCarthy.* Cambridge: MIT, 2001. vi, 456 p. Paper $25.00.

RICHARDSON, JOHN, and BRIAN LEITER, eds.: *Nietzsche.* New York: Oxford, 2001. viii, 379 p. Paper $19.95.

RILEY, PATRICK, ed.: *The Cambridge Companion to Rousseau.* New York: Cambridge, 2001. xii, 453 p. Cloth $69.95.

ROBBINS, JILL, ed.: *Is it Righteous To Be? Interviews with Emmanuel Levinas.* Stanford: University Press, 2001. viii, 305 p. Cloth $55.00, paper $24.95.

RORTY, AMÉLIE OKSENBERG, ed.: *The Many Faces of Evil: Historical Perspectives.* New York: Routledge, 2001. xviii, 346 p. Cloth $100.00, paper $29.95.

ROWE, WILLIAM L., ed.: *God and the Problem of Evil.* Malden, MA: Blackwell, 2001. xiii, 317 p. Cloth $64.95, paper $24.95.

RUSSELL, BERTRAND: *Yours Faithfully, Bertrand Russell: A Lifelong Fight for Peace, Justice, and Truth in Letters to the Editor.* RAY PERKINS, JR., ed. Chicago: Open Court, 2001. xii, 433 p. Cloth $52.95, paper $26.95.

SCOTT, CHARLES E., SUSAN M. SCHOENBOHM, DANIELA VALLEGA-NEU, and ALEJANDRO VALLEGA, eds.: *Companion to Heidegger's* Contributions to Philosophy. Bloomington: Indiana UP, 2001. 255 p. Cloth $49.95, paper $22.95.

SEIGFRIED, CHARLENE HADDOCK, ed.: *Feminist Interpretations of John Dewey.* University Park: Penn State UP, 2002. xi, 317 p. Cloth $65.00, paper $24.50.

SIBLEY, FRANK: *Approach to Aesthetics: Collected Papers on Philosophical Aesthetics.* JOHN BENSON, BETTY REDFERN, and JEREMY ROXBEE COX, eds. New York: Oxford, 2001. xi, 280 p. Cloth $45.00.

SIMON, ROBERT L., ed.: *The Blackwell Guide to Social and Political Philosophy.* Malden, MA: Blackwell, 2002. xi, 329 p. Paper $34.95.

SINGER, PETER: *Unsanctifying Human Life: Essays on Ethics.* HELGA KUHSE, ed. Malden MA: Blackwell, 2002. vii, 399 p. Cloth $59.95.

STAINTON, ROBERT J., ed.: *Perspectives in the Philosophy of Language: A Concise Anthology.* Ontario: Broadview, 2000. 352 p. Paper $22.95.

STEWARD, M.A., ed.: *English Philosophy in the Age of Locke.* New York: Oxford, 2001. x, 326 p. Cloth $60.00.

TESSMAN, LISA, and BAT-AMI BAR ON, eds.: *Jewish Locations: Transversing Racialized Landscapes.* Lanham, MD: Rowman and Littlefield, 2001. 251 p. Cloth $70.00, Paper $22.95.

TOWNSEND, DABNEY, ed.: *Eighteenth Century British Aesthetics.* Amityville, NY: Baywood, 1999. viii, 486 p. Cloth $60.00.

TRIGO, BENIGNO, ed.: *Foucault and Latin America: Appropriations and Deployments of Discoursive Analysis.* New York: Routledge, 2002. xxi, 305 p. Paper $24.95.

TUANA, NANCY, and SANDRA MORGAN, eds.: *Engendering Rationalities.* Albany: SUNY, 2001. xii, 413 p. Cloth $75.50, paper $25.95.

VAN GERWEN, ROB, ed.: *Richard Wollheim on the Art of Painting: Art as Representation and Expression.* New York: Cambridge, 2001. viii, 285 p. Cloth $55.00.

VAN NORDEN, BRYAN W., ed.: *Confucius and the Analects: New Essays.* New York: Oxford, 2002. x, 342 p. Paper $24.95.

WARD, GRAHAM, ed: *The Blackwell Companion To Postmodern Theology.* Malden, MA: Blackwell, 2001. xii, 530 p. Cloth $124.95.

WYSCHOGROD, EDITH, JEAN-JOSEPH GOUX, and ERIC BOYNTON, eds.: *The Enigma of Gift and Sacrifice.* New York: Fordham, 2002. vi, 186 p. Cloth $35.00, paper $20.00.

TAKE YOUR PHILOSOPHY DEGREE INTO THE CLASSROOM!

Thousands of children and adolescents around the world do philosophy as a regular part of their curriculum AND THEY LOVE IT!

At Montclair State University's **Institute for the Advancement of Philosophy for Children*** you can learn how to practice philosophy with children as a classroom teacher and/or to work with teachers who want philosophy in their classrooms.

Choose from three degree programs:

- Doctorate of Education (Ed D.), Specialization in Philosophy for Children
- Masters of Education (M. Ed.), Concentration in Philosophy for Children
- Master of Arts in Teaching (M.A T.), Concentration in Philosophy for Children (provides teacher certification.)

Assistantships are available!

To learn more about these innovative programs contact
Professor Megan Laverty at (973) 655-4351, LavertyM@mail.montclair.edu

To learn more about Philosophy for Children,
visit the IAPC web site at http://www.montclair.edu/IAPC

*Recipient of the American Philosophical Association/Philosophical Documentation Center's 2001 Award for Excellence and Innovation in Philosophy Programs

MONTCLAIR
STATE
UNIVERSITY

THE JOURNAL OF
PHILOSOPHY

The Journal is pleased to announce to our readers and subscribers the launch of its new website, which will have the pages listed below. The address is: *www.journalofphilosophy.org* or simply *www.jphil.org*.

~•~

PURPOSE & HISTORY
CONTACTS AND ADDRESSES
BOARD OF TRUSTEES
EDITORIAL BOARD
MEMOS TO AUTHORS
SUBSCRIPTIONS & ADVERTISING
RECENT & FORTHCOMING ISSUES
ELECTRONIC ISSUES
INDEX 1904–PRESENT
RIGHTS AND PERMISSIONS
LECTURES & CONFERENCES
PHILOSOPHY LINKS

~•~

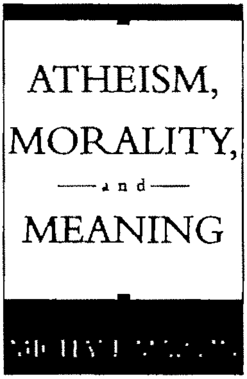
New Philosophy Titles from...

Prometheus Books

ATHEISM, MORALITY, AND MEANING
Michael Martin
235 pp • ISBN 1-57392-987-5 • PB $21

TOWARDS THE UNDERSTANDING OF KARL MARX
A Revolutionary Interpretation
Sidney Hook
Expanded Edition
Edited with an Introduction by Ernst B. Hook
Historical Introduction by Christopher Phelps
and Contributions by Lewis S. Feuer
and Paul Berman
465 pp • ISBN 1-57392-882-8 • HC $32

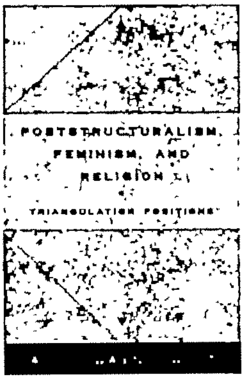
CONTROLLING TECHNOLOGY
Contemporary Issues
Second Edition
Edited by Eric Katz, Andrew Light, and William Thompson
550 pp • ISBN 1-57392-983-2 • PB $34



ATHEISM,
MORALITY,
and
MEANING

...and

Humanity Books



POSTSTRUCTURALISM,
FEMINISM, AND
RELIGION
TRIANGULATING POSITIONS

POSTSTRUCTURALISM, FEMINISM, AND RELIGION
Triangulating Positions
Carol Wayne White
184 pp • ISBN 1-57392-630-2 • HC $49

MIND, BODY, AND FREEDOM
Patrick T. Mackenzie
140 pp • ISBN 1-59102-001-8 • HC $45

SOCIALISM
Key Concepts in Critical Theory Series
Edited by Michael W. Howard
408 pp • ISBN 1-57392-956-5 • PB $29

[illegible] prometheusbooks.com
Call toll free [illegible] • www.prometheusbooks.com

PHILOSOPHY OF SCIENCE

The Official Journal
of the Philosophy
of Science
Association

Editor-in-Chief
Noretta Koertge
Indiana University

Associate Editors
W. Michael Dickson
Michael Friedman
Elisabeth A. Lloyd

Published quarterly, plus supplemental volume of the Proceedings of the Biennial Meeting of the Philosophy of Science Association.

Purpose and scope

Founded in 1934, ***Philosophy of Science*** has become the authoritative publication concerned with philosophical inquiry into the theory and practice of natural and social science. Adhering to the strictest standards in its selection of essays, discussion articles and book reviews, the ***Journal*** brings together the best scholarship on the methodological, ontological, epistemological, anthropological, and ethical premises and implications of the various scientific disciplines. In addition to new developments in traditional areas, ***Philosophy of Science*** presents ground-breaking work on conceptual issues that arise in new areas of scientific inquiry, such as bioinformatics, cognitive ethology, and neuroscience, as well as discussions of policy issues connected with the changing role of science in our society.

Intended Audience

Philosophy of Science is not just for philosophers! It will be of interest to natural scientists searching for new perspectives on their fields, epistemologists, and cognitive scientists who study the organization of knowledge, and social scientists and humanists who approach science as a cultural phenomenon.

Membership in the Philosophy of Science Association includes the journal ***Philosophy of Science*** *at no additional charge. Individual subscriptions to the* journal *and membership in the PSA are concurrent*

Year 2001 PSA membership dues/subscription rates: Institutions-A (Australia, Canada, Hong Kong, Japan, New Zealand, Western Europe, and U.S.) $102 00, Institutions-B (all other countries) $51.00. Members (income less than $40,00) $57 00; Members (income $40,000-$49,999) $67 00, Members (income $50,000-$59,999) $77 00, Members (income more than $60,000) $87 00, Members-B (all other countries) $25 00; Student members (with copy of valid ID) $25.00. Canadians, please add 7% GST plus postage. Institutions outside USA, please add $6 postage. MasterCard and Visa orders accepted (please include account number, expiration date, signature, and phone number), credit card orders may be sent via fax to (773) 753-0811or (877) 705-1879 Send your order with check, credit card information, or purchase order to

The University of Chicago Press
Journals Division, P.O. Box 37005, Chicago, IL 60637
Order online at www.journals uchicago.edu/PHILSCI

2/02

New from Princeton

Descartes's Method of Doubt
Janet Broughton

Descartes thought that we could achieve absolute certainty by starting with radical doubt. But why did Descartes think we should take these exaggerated doubts seriously? And if we do take them seriously, how did he think any of our beliefs could ever escape them? Janet Broughton seeks to answer these questions and to present a fresh way of understanding precisely what Descartes was up to.

"This stunning work is without question a major contribution...original, provocative, and philosophically interesting.... It is a brilliant piece of philosophy, a masterful work of scholarship, and an engaging read—a trio not often encountered."—Edwin McCann, University of Southern California
Cloth $26.95 ISBN 0-691-08818-7

W.V. Quine
Alex Orenstein

Willard Van Orman Quine (1908–2000) changed the way we think about language and its relation to the world. This introduction to Quine's philosophical ideas provides readers with an authoritative analysis of Quine's lasting contributions.

"This is an outstanding account of Quine's thought, and will be invaluable to anyone seeking a really good, accessible and insightful introduction to it."
—A. C. Grayling, Birkbeck College, University of London
Philosophy Now Series, John Shand, series editor
Paper $17.95 ISBN 0-691-09606-6
Cloth $55.00 ISBN 0-691-0960508 Due April
Available from Princeton in the Americas

Princeton University Press

Best in Philosophy from Cambridge

Theoretical Philosophy after 1781
Henry Allison and Peter Heath, Editors
The Cambridge Edition of the Works of Immanuel Kant in Translation
0-521-46097-2 Hardback $95.00

The Ethical Dimensions of the Biological and Health Sciences
A Textbook with Case Studies
Second Edition
Ruth Ellen Bulger, Elizabeth Heitman, and Stanley Joe Reiser, Editors
0-521-81053-1 Hardback $90.00
0-521-00886-7 Paperback $30.00

Jacques Derrida and the Humanities
A Critical Reader
Tom Cohen
0-521-62370-7 Hardback $70.00
0-521-62565-3 Paperback $25.00

Religious Conviction in Liberal Politics
Christopher J. Eberle
0-521-81224-0 Hardback $75.00
0-521-01155-8 Paperback $28.00

The Genealogy of Aesthetics
Ekbert Faas
0-521-81182-1 Hardback $65.00

Toleration as Recognition
Anna Elisabetta Galeotti
0-521-80676-3 Hardback $60.00

Liberal Pluralism
The Implications of Value Pluralism for Political Theory and Practice
William A. Galston
0-521-81304-2 Hardback $55.00
0-521-01249-X Paperback $19.00

Descartes' System of Natural Philosophy
Stephen Gaukroger
0-521-80897-9 Hardback $60.00
0-521-00525-6 Paperback $22.00

Karl Popper—The Formative Years, 1902-1945
Politics and Philosophy in Interwar Vienna
Malachi Haim Hacohen
0-521-89055-1 Paperback $25.00

Critique of the Power of Judgment
Immanuel Kant
Paul Guyer and Eric Matthews, Editors
The Cambridge Edition of the Works of Immanuel Kant in Translation
0-521-34892-7 Paperback $25.00

An Introduction to Classical Islamic Philosophy
Second Edition
Oliver Leaman
0-521-79343-2 Hardback $65.00
0-521-79757-8 Paperback $23.00

Symbols of Jesus
A Christology of Symbolic Engagement
Robert Cummings Neville
0-521-80787-5 Hardback $65.00
0-521-00353-9 Paperback $23.00

Moral Knowledge
Ellen Frankel Paul, Fred D. Miller, Jr., and Jeffrey Paul, Editors
Social Philosophy and Policy 18 2
0-521-00603-1 Paperback $25.00

Consciousness and Language
John R. Searle
0-521-59237-2 Hardback $65.00
0-521-59744-7 Paperback $23.00

Now in paperback...

Descartes and Augustine
Stephen Menn
0-521-01284-8 Paperback $25.00

Available in bookstores or from
CAMBRIDGE UNIVERSITY PRESS
800-872-7423
us.cambridge.org/philosophy

Philosophy from Yale

Gadamer in Conversation

Reflections and Commentary

Hans-Georg Gadamer

Edited and translated by Richard E. Palmer

This volume presents six lively conversations in which Gadamer takes up key issues in his philosophy, addresses points of controversy, and replies to his critics, including those who accuse him of having been in complicity with the Nazis. Yale Studies in Hermeneutics $22.50

Peiresc's Europe

Learning and Virtue in the Seventeenth Century

Peter N. Miller

Winner of the 2000 Jacques Barzun Prize in Cultural History sponsored by the American Philosophical Society

20 illus $40.00

Summulae de dialectica

John Buridan

An annotated translation, with a philosophical introduction by Gyula Klima

This volume, the first annotated translation of the entire text is "an **excellent translation of an important work**."—E. J. Ashworth

Yale Library of Medieval Philosophy Series

12 illus $120.00

Yale University Press

yalebooks.com

The Empirical Stance

Bas C. van Fraassen

"**This work argues powerfully** for a radical redirection of much contemporary thinking in epistemology and other important topics. It should force all of us to rethink the nature of analytic philosophy."—Paul Teller $30.00

Available in Paper

The Social Contract and *The First and Second Discourses*

Jean-Jacques Rousseau

Edited and with an Introduction by Susan Dunn

With essays by Gita May, Robert N. Bellah, David Bromwich, and Conor Cruise O'Brien

Accompanying essays by major scholars convey a sense of the still vital and contentious debate that surrounds Rousseau, his ideas, and his legacy. Rethinking the Western Tradition

$35.00 cloth; $15.95 paperback

The Patient as Person

Exploration in Medical Ethics, 2nd Edition

Paul Ramsey, with a new foreword by Margaret A. Farley and essays by Albert R. Jonsen and William F. May

"A **masterpiece of thoroughness** in evaluating conflicting moral claims which become explicit in crucial medical situations."—Dolores Dooley-Clarke, *Philosophical Studies* (on the earlier edition) Published under the auspices of ISPS $17.95 paperback

THE JOURNAL OF PHILOSOPHY

UNIVERSITY
5/8/02

VOLUME XCIX, NUMBER 6
JUNE 2002

page
279 *The Ersatz Pluriverse* Theodore Sider
316 *Farewell to the Direct Argument* David Widerker

COMMENTS AND CRITICISM
325 *Stanley and Williamson on Knowing How* John Koethe

BOOK REVIEWS
329 Wolfgang-Rainer Mann: *The Discovery of Things: Aristotle's* Categories *and Their Context* James Allen

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721. You may also visit our website at: www.journalofphilosophy.org

THE JOURNAL OF PHILOSOPHY

2002

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2002 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME XCIX, NO. 6, JUNE 2002

9\18\02

THE ERSATZ PLURIVERSE*

Although many are impressed with the utility of possible worlds in linguistics and philosophy, few can accept the modal realism of David Lewis,[1] who regards possible worlds as *sui generis* entities of a kind with the concrete world we inhabit.

Not all uses of possible worlds require exotic ontology. Consider, for instance, the use of Kripke models to establish formal results in modal logic. These models contain sets often regarded for heuristic reasons as sets of "possible worlds." But the "worlds" in these sets can be anything at all; they can be numbers, or people, or fish. The set of worlds, together with the accessibility relation and the rest of the model, is used as a purely formal structure.[2] One can even go beyond establishing results about formal systems and apply Kripke models to English, as Charles Chihara[3] has recently argued. He shows, for

---

* Since writing this article, I have learned that Daniel Nolan has independently developed a somewhat similar account; see footnote 11. Previous versions were presented at the University of Rochester, the 1997 American Philosophical Association Pacific Division meeting, the 1997 Notre Dame Mighty Midwestern Metaphysician's conference, Syracuse University, Princeton University, and Yale University. I would like to thank David Braun, Phillip Bricker, Cian Dorr, Tamar Szabó Gendler, Jeff Goodman, John Hawthorne, Mark Johnston, Tom McKay, John Mouracade, Mark Moyer, Brent Mundy, Daniel Nolan, Alvin Plantinga, Gideon Rosen, Ed Wierenga, and, especially, Kit Fine, for helpful comments. The main idea behind the pluriverse theory occurred to me while thinking about an early draft of Joseph Melia's "Reducing Possibilities to Language" (*Analysis*, LXI (2001): 19-29), although he is not responsible for shortcomings of the present article. Finally, I am grateful to David Lewis for his comments, and for his writings throughout the years, without which this article and many others like it would not exist.

[1] *On the Plurality of Worlds* (New York: Blackwell, 1986).

[2] See G. E. Hughes and M. J. Cresswell, *A New Introduction to Modal Logic* (New York: Routledge, 1996).

[3] *The Worlds of Possibility* (New York: Oxford, 1998).

0022-362X/02/9906/279–315

instance, how to use Kripke models (plus primitive modal notions) to give an account of validity for English modal sentences. In other cases, worlds are not really needed at all. It is often vivid to give a counterexample thus: "There is a possible world in which *P*. Since your theory implies that in all worlds, not-*P*, your theory is wrong." But the counterexample could just as easily be given using modal operators: "Possibly, *P*. Since your theory implies that it is necessary that not-*P*, your theory is wrong."

Unfortunately for those who yearn for desert landscapes, many interesting applications require more ontological seriousness. In *The Conscious Mind*, David Chalmers[4] uses worlds to set up his two-dimensional modal framework. In defining the primary intension of 'water' as a function that assigns to any (centered) world, *w*, the class of things that would count as water in *w* when *w* is "considered as actual," Chalmers is not thinking of the set of worlds as a purely mathematical structure. He means to refer to a *particular* function defined on a space of genuine possible worlds.

Or consider Lewis's[5] formulation of materialism as the claim that no two possible worlds lacking "alien natural properties" differ without differing physically. This is no claim about the formal structure of Kripke models; it is a claim about a particular class of worlds, the class of worlds that lack alien natural properties.

Other examples abound. Linguists and philosophers of language utilize set-theoretic constructions out of possible worlds and individuals as semantic values. Philosophers use possible worlds to define probability functions, properties, and propositions. Worlds are used to state theories, make claims, formulate supervenience theses, and illuminate distinctions. Worlds may be used to give the truth conditions for ordinary quantification over possibilities: for example, 'There are at least five ways to win this chess match'. The journals and books published in the last forty years contain hundreds of uses of possible worlds, few of which could be reconstructed as purely formal uses of Kripke models or as Chihara-style applications. Moreover, these invocations of possible-worlds talk usually go far beyond using possible-worlds language as a kind of vivid shorthand for sentences containing modal operators: it is well known that a language employing quantification over possible worlds and individuals has more expressive power than the language of modal predicate logic.[6]

---

[4] New York: Oxford, 1996; chapter 2.

[5] "New Work for a Theory of Universals," *Australasian Journal of Philosophy*, LXI (1983): 343-77; especially pp 361-65.

Possible-worlds semantics and metaphysics appear to require genuine worlds. And yet, who can believe Lewis's modal realism?

Like many, I turn to reduction. An adequate reduction of talk of possibilia—worlds and their inhabitants—must be materially adequate, but must not appeal to objectionable entities (like Lewisian worlds). Given such a reduction, the contemporary possible-worlds theorist's tools may be freely used with a clear ontological conscience.

A good starting point is what Lewis calls "linguistic ersatzism," which identifies possible worlds with maximal consistent sets of sentences. But linguistic ersatzism faces an apparently devastating problem: the problem of descriptive power. The solution to this problem will lead to an unfamiliar but attractive theory.

Because of my own views (which I shall not defend here), the theory to be developed will assume world-bound individuals and counterpart theory.[7] But the theory can be developed under other assumptions.

Section I describes the problem of descriptive power for linguistic ersatzism. Section II introduces the theory to be defended and describes its solution to the problem of descriptive power. Section III then gives an extensive formal development of the theory. Section IV replies to objections, and section V compares the theory with modal fictionalism.

I. LINGUISTIC ERSATZISM AND THE PROBLEM OF DESCRIPTIVE POWER

The story is a familiar one. We find ourselves apparently quantifying over things we regard as ontologically objectionable. We therefore reinterpret this quantification as *really* being over things more easily accepted. The instance of the story at hand is the reduction of possible worlds to linguistic-ersatz worlds.

A linguistic-ersatz possible world is a maximal consistent set of sentences. To construct a possible world in which a donkey talks, we need only include the sentence 'A donkey talks' in a set, along with enough other sentences to ensure that the set is not silent about any matter (maximality), but not so many that sentences in the set contradict each other (consistency). The notion of consistency here must be modal: a set is consistent if and only if it is possible that all the members of the set be true together. For mere logical consistency will not do: sets of sentences asserting the existence of married bachelors and round squares may be logically consistent but will not correspond to possible worlds. Unless modal consistency can be reduced in some way, linguistic-ersatz worlds cannot be used in a reductive analysis of modality, on pain of circularity. But the linguistic ersatzist can accept this limitation. The reduction of worlds to language still has a point, for it allows us to reduce all talk of worlds—which runs far beyond that which can be said utilizing merely the modal operators—to talk of possibility and necessity. As for these, they may one day be reduced in some way that does not involve worlds, or they may remain primitive.[8]

---

[6] See Allen Hazen, "Expressive Completeness in Modal Languages," *Journal of Philosophical Logic*, V (1976): 25-46, especially section 3. Even if one enriches modal predicate logic in various ways—for example with an actuality operator, or even the bolder enhancements of Graeme Forbes, *The Metaphysics of Modality* (New York: Oxford, 1985), pp. 90-93—it is far from clear how to simulate full quantification over set-theoretic constructions of possibilia.

[7] See Lewis, "Counterpart Theory and Quantified Modal Logic," this JOURNAL, LXV, 5 (March 7, 1968): 113-26. I defend counterpart theory (both modal and temporal) in my *Four-Dimensionalism: An Ontology of Persistence and Time* (New York: Oxford, 2001).

Certain objections are quickly answered by taking a broad view of what counts in the present context as a sentence. Let us allow infinitely long sentences;[9] and let the language of those sentences be *Lagadonian*, in that objects, properties, and relations count as names of themselves, so that every actually existing object, property, or relation has a name.[10] Natural languages are neither infinitary nor Lagadonian, but there is no need to take 'sentence' or 'language' very seriously. The world-making language need not be learnable or speakable (though the difficulty of writing Lagadonian sentences will require me to revert to English in examples). All that is needed is that its "sentences" have well-defined meanings and be ontologically unobjectionable. These sentences can be understood as mathematical objects—infinite sequences of Lagadonian names, quantifiers, variables, and so on.

There are various ways to fill in this sketch of linguistic ersatzism (*ibid.*, chapter 3.2), but all face the following *problem of descriptive power*.[11] An actual language apparently cannot fully describe possibilities involving things that do not actually exist. The ersatzist can attempt to construct these possibilities using qualitative sentences describing what nonactual entities would be like, but such attempts conflate distinct possibilities. The problem comes in two forms, one involving nonactual individuals, the other involving nonactual fundamental properties and relations.

---

[8] See my "Reductive Theories of Modality," in Michael J. Loux and Dean W. Zimmerman, eds., *The Oxford Handbook of Metaphysics* (New York: Oxford, forthcoming).

[9] See Phillip Bricker, "Reducing Possible Worlds to Language," *Philosophical Studies*, LII (1987): 331-55, especially pp 343-49; and M. A. Dickmann, *Large Infinitary Languages: Model Theory* (Amsterdam: North-Holland, 1975).

[10] See Lewis, *On the Plurality of Worlds*, pp. 145-46.

[11] See Lewis, *On the Plurality of Worlds*, pp. 158-65; Bricker, pp. 349-53; Alan McMichael, "A Problem for Actualism about Possible Worlds," *Philosophical Review* XCII (1983): 49-66. For excellent survey and discussion, see Hazen, "Actualism Again," *Philosophical Studies*, LXXXIV (1996): 155-81, and Daniel Nolan, *Topics in the Philosophy of Possible Worlds* (New York: Routledge, 2002), chapter 5.

A possible world in which I am six feet tall may be constructed by embedding the sentence 'Ted is six feet tall' in a maximal consistent set of sentences. But one cannot thus construct an ersatz world in which some individual that does not actually exist is six feet tall, since we have no names for nonactual individuals. Instead, one must use a descriptive sentence:

> There is some individual distinct from Ted, Tom, Dick, Harry,..., who is 6 feet tall.

where the list 'Ted, Tom, Dick, Harry,...' includes a name for each actual individual. Worlds with nonactual individuals are thereby constructed from *actualist* descriptions of what those individuals would be like—descriptions that name only actual individuals and qualitative, general characteristics. But this procedure will never distinguish between worlds that differ solely over which nonactual individuals play which qualitative roles. We could construct a world in which one man wears a mask of a bat, drives a car, and fights crime with a younger masked chum named 'Robin'; but what of a distinct world in which the person who in the first world wears the bat-mask is now named 'Robin', wears the Robin mask, and so on, and who is replaced in the Batman role by the person who wore the Robin mask in the first world? No sentences in an actual world-making language will distinguish these worlds, so the current proposal identifies the apparently distinct worlds with a single set of sentences, and hence with each other.

The objection assumes there can be qualitatively identical worlds differing in which qualitative roles are played by which individuals. This assumption is a controversial doctrine sometimes called *haecceitism*. Its opponents, antihaecceitists, say that worlds vary only in their qualitative descriptions; therefore the limitation of the linguist ersatzist to actualist descriptions is no limitation at all. Given my acceptance of counterpart theory, I accept antihaecceitism and thus do not mind this limitation; but it is better to have an account of worlds that is independent of this controversial doctrine. There is, moreover, an analog to the objection that does not depend on haecceitism.

Possibilities involving nonactual *properties*, as well as individuals, must be accounted for. As before, the ersatzer can use descriptions of the roles such properties would play; but as before, such descriptions will not distinguish possible worlds in which nonactual properties swap roles.

There could have existed two fundamental types of matter, call them A-matter and B-matter, which do not actually exist, playing a certain nomic role which may be partially described as follows: A-matter attracts both negatively and positively charged things, whereas B-matter repels each. Let us understand 'fundamental' so that neither supervenes on, or may be constructed in any way from, properties and relations instantiated in the actual world. How can we construct an ersatz world corresponding to this possibility? We have no names for the properties since they do not in fact exist. Our best attempt will be to use descriptions, perhaps describing the roles in the laws of nature that A- and B-matter would play, or perhaps describing the pattern of distribution throughout spacetime A- and B-matter would have. Choosing the former course, let us embed the following sentence in a maximal consistent set:

> (1) There are two fundamental properties, *P* and *Q*, such that (i) neither *P* nor *Q* is identical to charge, charm,..., and (ii) it is a law of nature that objects with *P* attract both negatively charged things and positively charged things, while objects with *Q* repel negatively and positively charged things.

But we can also imagine a distinct world in which A-matter and B-matter have swapped nomic roles: in this new world it is B-matter rather than A-matter that attracts charged particles. Since neither sentence (1) nor any other actual sentence will distinguish these worlds, linguistic ersatzism fails as a general reduction of possible-worlds talk.

Some will reply that since properties are abstract objects, they exist necessarily (or better, that necessarily, every property exists necessarily). If the reply were correct, then uninstantiated properties of *being made up of A-matter* and *being made up of B-matter* would exist in the actual world, could serve as Lagadonian names of themselves, and could be included in sentences that would distinguish the worlds in question. One could construct, for example, a sentence saying that objects made up of A-matter attract charged things, which would be true of the first world but not the second.[12] But I cannot accept the reply, for two reasons.[13] First, it requires an ontology on which the existence of properties is radically independent of the goings-on of the concrete world. The theory of worlds should not be thus held hostage to the theory of properties. The properties required for the reply could not be sets of their instances (each would need to be the empty set; A-matter and B-matter would then be identified with each other). Nor could they be immanent universals (in the sense of D. M. Armstrong[14]) which are supposed to be "wholly present in their instances" and incapable of existing uninstantiated. Even Michael Tooley,[15] whose "transcendent" universals can exist uninstantiated in certain cases, would not accept uninstantiated universals of A-matter and B-matter. For Tooley, uninstantiated universals are accepted only when they play a role in the laws of nature, and we may stipulate that A-matter and B-matter play no role in the actual laws of nature.

---

[12] An analogous reply may be made to the objects version of the problem of descriptive power, based on uninstantiated individual essences, as in section 5 of Alvin Plantinga, "Actualism and Possible Worlds," *Theoria*, XLII (1976): 139-60. The criticisms in the text apply to this reply as well; but there are additional problems:

Note further that since the properties of being made up of A-matter and B-matter would be fundamental, they could not be constructed in any way, even by infinitary means, from less problematic properties, on any of the theories of properties just mentioned.

The second problem with the reply is its ontological extravagance. One of the main reasons to reduce worlds is parsimony. As W. V. Quine[16] puts it colorfully, the believer in possibles accepts an "overpopulated" "bloated universe" which is "in many ways unlovely," and "offends the aesthetic sense of us who have a taste for desert landscapes" (*ibid.*, pp. 3, 4). If nonactualized possibles bloat the universe, so would the actual but uninstantiated properties needed for the reply.

Neither reason for resisting uninstantiated fundamental properties is a reason for rejecting uninstantiated properties as such. On some views, complex properties are composed of or constructed from others, so that constructed uninstantiated properties would be no more objectionable than the properties from which they are con-

---

see section 2 of Robert Merrihew Adams, "Actualism and Thisness," *Synthese*, XLIX (1981): 3-41, and Kit Fine, "Plantinga on the Reduction of Possibilist Discourse," in J Tomberlin and P. van Inwagen, eds , *Alvin Plantinga* (Boston: Reidel, 1985), pp. 145-86, especially sections 2 and 3.

[13] See also McMichael, pp. 59-61; and Lewis, *On the Plurality of Worlds*, pp. 160-61.

[14] *A World of States of Affairs* (New York: Cambridge, 1997); *Universals: An Opinionated Introduction* (Boulder: Westview, 1989); *Universals and Scientific Realism, Volumes I and II* (New York: Cambridge, 1978).

[15] *Causation. A Realist Approach* (New York: Oxford, 1987), especially section 3.1.4, pp. 119-20.

[16] "On What There Is," in his *From a Logical Point of View* (Cambridge: Harvard, 1953), pp. 1-19.

structed. My quarrel is with uninstantiated *fundamental* properties, which cannot be constructed from less problematic ones.

Just as the objects-version of the problem of descriptive power can be avoided by accepting a metaphysical thesis about individuals, namely, antihaecceitism, the properties-version of the problem may be avoided by going in for certain metaphysical claims about properties. For example, if Sydney Shoemaker and Chris Swoyer[17] are right that properties have their nomic roles essentially, then the problematic case of nomic role swapping is impossible. Alternatively, one could claim that the identity of a property is determined in some way by its pattern of instantiation. One could then safely account for alien fundamental properties without fear of conflating possibilities with descriptions of the patterns of instantiation such properties would have.[18] But a theory of worlds should be independent of these matters.

It seems that the linguistic ersatzer can solve the problem of descriptive power only by accepting both antihaecceitism and some strong claim about the identity conditions for properties.[19] This is an unhappy situation. Using possible-worlds talk should not commit one to controversial metaphysical doctrines.

II A NEW APPROACH TO REDUCING WORLDS

Fortunately, such commitments can be avoided. A theory of worlds may be constructed which distinguishes possibilities in which nonactual individuals or fundamental properties swap roles, without making special assumptions about the metaphysics of individuals or properties.

Return to the worlds where A-matter and B-matter swap nomic roles. We cannot distinguish the worlds from within, by giving them different descriptions. But we can distinguish the worlds from an external perspective, so to speak. We can say: "There are two distinct worlds and a pair of nonactual properties; in one world the pair instantiates certain nomic roles; in the other world, the nomic roles of the properties are reversed."

---

[17] Shoemaker, "Causality and Properties," in Peter van Inwagen, ed., *Time and Cause: Essays Presented to Richard Taylor* (Boston: Reidel, 1980), pp. 109-35; Swoyer, "The Nature of Natural Laws," *Australasian Journal of Philosophy*, LX (1982): 203-23. For an objection see Lewis, *On the Plurality of Worlds*, pp. 162-63.

[18] See Mark Heller, "Property Counterparts in Ersatz Worlds," this JOURNAL, XCV, 6 (June 1998): 293-316.

[19] My survey of solutions has not been completely exhaustive. In "Time and Thisness," Adams, a haecceitist, paraphrases the sentence (RS): "there exist two possible worlds in which a pair of non-actual individuals swap roles" as, roughly, the assertion that there exists a possible world, w, according to which the following is true: there exist two nonactual individuals, x and y, and two ersatz possible worlds, w' and w'', that are alike except for an exchange of roles by x and y (p. 36). But Adams paraphrases other sentences that quantify over worlds—for example, 'There exist two worlds, one in which a donkey talks, another in which a bluefish walks'—as genuine quantifications over ersatz worlds (for Adams, maximal consistent sets of propositions). A uniform treatment would be preferable.

The theory I propose departs radically from linguistic ersatzism. I do away with ersatz worlds and individuals in reducing worlds talk, and instead use a single *ersatz pluriverse*, a single abstract entity that represents the totality of possible worlds and individuals all at once. Two ways of carrying out this idea will be considered below; here I consider the version in which the abstract entity is a *pluriverse sentence*, which looks roughly as follows:

> THERE ARE worlds w_1, w_2,...and THERE ARE properties and relations P_1, P_2,...that are distinct from the following actual properties and relations:..., and THERE ARE possible individuals x_1, x_2,...that are distinct from the following actual individuals:..., SUCH THAT: ...w_1...and...w_2...and....

At the end of the existentially quantified pluriverse sentence, there are conjoined open formulas, one for each possible world. One of these "world conjuncts" might look like this: "x_1 is in w_1, and has property P_2, and...." The language of the pluriverse sentence must be an infinitary language, since there will need to be infinitely many existential quantifiers and infinitely many world conjuncts, and also some infinitely long world conjuncts.

This requires a departure from the usual way of reducing possible-worlds talk. Most familiar reductions are *entity-for-entity* reductions, in that they provide a surrogate entity for each entity to be reduced, an abstract surrogate for each possible world.[20] I propose instead a *holistic* reduction. Instead of individual world-surrogates, I offer a single pluriverse-surrogate. Worlds talk cannot, therefore, be talk about possible world-surrogates; it must instead be talk about the single surrogate. Suppose we want to say that there is a possible world in which a donkey talks. The usual reductionist paraphrase is that there is a possible world-surrogate of some sort according to which a donkey talks. (The linguistic ersatzist's surrogate is a maximal consistent set of sentences; other theories provide other surrogates.) On the pluriverse view, we say instead: according to the pluriverse sentence, there is a possible world in which a donkey talks. Thus, quantification over possibles is interpreted, not as quantification over surrogates, but rather as truth of a quantified sentence according to a single surrogate.

---

[20] See Adams, "Theories of Actuality," *Noûs*, VIII (1974): 211-31; Plantinga; and Robert Stalnaker, "Possible Worlds," *Noûs*, X (1976). 65-75. Fine (*op. cit.*, pp. 180-83); and Rosen ("Modal Fictionalism," *Mind*, XCIX (1990): 327-54) are exceptions.

When I speak of what is true *according to* the ersatz pluriverse, I am not proposing that we take 'according to' as a mysterious or novel primitive. (This is in contrast to Gideon Rosen's modal fictionalism, which I discuss below.) 'According to' just means logical entailment of a sort to be explained. (The entailment must not be strict implication, in the sense of the necessity of the material conditional, for it may well be impossible that the pluriverse sentence is true, in which case the pluriverse sentence would strictly imply every sentence.) Thus, a sentence *S* quantifying over possibilia is reinterpreted as the claim that the pluriverse sentence logically entails *S*.

A "strict and philosophical" interpretation of quantification over worlds must be distinguished from a more everyday interpretation. I use the words 'there are no possible worlds' to deny modal realism, and yet I quantify over worlds when such quantification is reduced via the ersatz pluriverse. The former denial is not intended to receive the ersatz-pluriverse analysis. What does get analyzed is quantification over possibilia by philosophers and linguists, as well as ordinary quantification over possibilities, as when a chess master remarks that there are at least five possibilities for winning a certain chess match.

This holistic approach to reducing worlds is attractive. The object of the reduction is the entire pluriverse, so it is natural to produce a surrogate for that entity. Moreover, the method of representation of this surrogate is parallel to that of linguistic ersatzism. A maximal consistent set of sentences represents a state of affairs if and only if that set contains sentences that entail that the state of affairs obtains. The pluriverse sentence also represents by entailment: it represents there being a certain sort of world or possible individual by (logically) entailing that such a world or individual exists.

Even before it is spelled out in detail, the pluriverse view can be seen to solve the problem of descriptive power. Assuming properties have neither their nomic roles nor their patterns of distribution essentially, any adequate theory of worlds should allow a pair of possible worlds in which two nonactual fundamental properties swap nomic roles. On the pluriverse view, the sentence 'There are such-and-such possible worlds' is reinterpreted as: 'The pluriverse sentence entails that there are such-and-such possible worlds'. Provided the pluriverse sentence is spelled out appropriately, it will logically entail the following sentence asserting the existence of the desired pair of worlds:

(*) There are two worlds, w_1 and w_2, and two nonactual, fundamental properties, P_1 and P_2, such that: P_1 plays nomic role R_1 in w_1 and P_2 plays nomic role R_2 in w_1, and P_1 plays nomic role R_2 in w_2 and P_2 plays nomic role R_1 in w_2.

Haecceitists could give a similar response to the problem of alien individuals—the pluriverse sentence could be spelled out so that it entails the sentence:

(**) There are two worlds, w_1 and w_2, and two nonactual individuals x_1 and x_2, such that: x_1 plays qualitative role R_1 in w_1 and x_2 plays qualitative role R_2 in w_1, and x_1 plays qualitative role R_2 in w_2 and x_2 plays qualitative role R_1 in w_2.

(Since I am no haecceitist, I choose not to spell out the pluriverse sentence in this way; rather, I reject the existence of worlds differing only by what individuals play what qualitative roles.)

The linguistic ersatzer had to represent nonactual individuals and properties by including existentially quantified sentences within ersatz worlds:

$$\overset{w_1}{\boxed{\exists P\,\phi(P)}} \qquad\qquad \overset{w_2}{\boxed{\exists Q\,\psi(Q)}}$$

But then one ersatz world could not make any assertions about a particular nonactual individual represented within another, since the variables for nonactual individuals were bound to quantifiers that occurred only *within* ersatz worlds. The solution is to move the existential quantifiers outside of the individual-world surrogates, to bind occurrences of variables within multiple-world surrogates:

$$\exists P \exists Q \left[\overset{w_1}{\boxed{\phi(P,Q)}} \qquad\qquad \overset{w_2}{\boxed{\psi(P,Q)}} \right]$$

Provided that the details of the pluriverse theory can be adequately filled in, the problem of descriptive power has been solved. The next section concerns those details.

I shall develop the pluriverse view within a broadly linguistic-ersatzist framework, but my solution to the problem of descriptive power is quite general. Other theories face the problem, and can benefit from the pluriverse strategy. Robert Merrihew Adams, Alvin Plantinga, or Robert Stalnaker could convert their nonlinguistic abstract possible worlds into a single nonlinguistic abstract pluriverse;

likewise, Armstrong could incorporate my methods into his combinatorial theory.[21]

III. THE CONSTRUCTION OF THE ERSATZ PLURIVERSE

Here, I provide a rigorous construction of the ersatz pluriverse. In outline, the construction will run as follows. I shall introduce *modal models*, which are similar to Kripke models. Modal models will be called *realistic* when they are faithful to the modal facts. The modal facts will be stated in a *modal language* (containing the $\Box$ and $\Diamond$); truths in this language will determine which modal models are realistic. A second language, the *possibilist language*, contains the sentences to be reduced—sentences about possibilia. The proposed reduction comes in two versions. The version alluded to in the previous section utilizes pluriverse sentences, which are constructed in the possibilist language as maximal descriptions of realistic models. At the end, I introduce a nonlinguistic version that uses realistic models directly: a sentence, *S*, in the possibilist language is reinterpreted as meaning that *S* is true in all realistic models.

IIIa. Two languages. I begin with a specification of a *possibilist language*, a language with the resources to describe possible worlds and individuals. This language is in essence the world-making language used by the linguistic ersatzer. Pluriverse sentences will be sentences of this language. Additionally, the possibilist language contains the target sentences of the reduction, those sentences about possibilia which are to be analyzed.

Lewis distinguishes between "rich" and "poor" world-making languages. A poor language names only a select few properties and relations, perhaps those of fundamental physics, and therefore will remain silent about many matters. I shall use instead a rich language, a language with a Lagadonian name for every (actual, concrete) individual, and a Lagadonian name for every (actual) property and relation (from now on, just "property").[22] The Lagadonian names of properties should include higher-order properties of properties (so that sentences like (*) from the previous section may be formulated). Note the use of properties as Lagadonian *names* rather than *predicates*.

---

[21] See Adams, "Theories of Actuality"; Plantinga; Stalnaker; Armstrong, *A Combinatorial Theory of Possibility* (New York: Cambridge, 1989).

[22] 'Concrete' rules out at least sets, and any other objects that would introduce cardinality problems. Likewise, the properties and relations should be restricted to properties and relations of nonsets. I shall speak as if the Lagadonian name of x is just x itself, but officially let it be $\langle 0,x\rangle$. Then we can distinguish between an occurrence of an entity as a variable and an occurrence of a Lagadonian name for that variable. I presuppose an "abundant" construal of properties (see Lewis, *On the Plurality of Worlds*, pp. 59-69).

Given this, the language must also include an instantiation predicate. It will be convenient to make the instantiation predicate multigrade; thus, 'x instantiates P' and 'x and y instantiate R' are both well formed. The sentences of this language are just sequences of the bits of primitive vocabulary: Lagadonian names, the instantiation predicate, quantifiers, Boolean connectives, and so on. The pluriverse theory thus requires set theory, plus the existence of properties and relations (which are the Lagadonian names of themselves)— defending nominalism is not my concern here. The language is infinitary both in allowing infinite sentences and also infinite blocks of quantifiers, in each case of arbitrarily long infinite length (I discuss cardinality worries in section IVA below.) The language contains no modal operators. Instead, it contains resources to speak explicitly about worlds: a syntactically distinct category of world-variables w_i, an individual constant $@$ for the actual world, and a two-place predicate 'exists in'.

Quantification in the possibilist language is first order. But since the pluriverse theory is to be combined with counterpart theory, there must be certain differences between quantification over individuals and over properties, since the ersatz pluriverse must represent the former but not the latter as being world-bound. Variables for properties are syntactically distinct from variables for individuals; and individual constants for (actual) individuals are syntactically distinct from constants naming (actual) properties.[23] Since individuals are to be represented as world-bound, the instantiation of properties by individuals may be represented as being instantiation *simpliciter*. But the properties will be represented as recurring in different worlds (as in sentence (*) above); hence the ersatz pluriverse must represent higher-order properties as being instantiated *at worlds*. Thus, a distinct predicate for higher-order instantiation must be introduced, multigrade as before, but with an extra place for worlds "$P_1,\ldots, P_n$ *instantiate* Q *at* w." Within the class of property constants and variables, let there also be a syntactic distinction between those for properties of individuals (type 0), and higher-order (type 1) properties of properties; and let there be a syntactically associated number of places with each property constant, namely, the number of places of that Lagadonian constant itself. (Syntactically, the constants are names, not predicates, so the number of places does not affect grammar; rather, it will constrain the interpretation of these constants within models.)

---

[23] Since the constants are Lagadonian, this is no stipulative claim about the shapes of the constants! It simply amounts to differential treatment of different sorts of constants at various points in the definitions that will follow.

The second language is the *modal language*. This language differs from the possibilist language by dropping the world variables, @, and 'exists in', and including in their place the modal operators □ and ◇ and a sentential actuality operator. Moreover, the predicate for higher-order instantiation in the modal language no longer has an argument place for worlds. Otherwise, the modal language is like the possibilist language: it is an infinitary first-order language with Lagadonian names for all actual concrete individuals and properties, and syntactic distinctions between constants and variables for individuals and properties.

IIIb. Modal models. My modal models are familiar: they are based on a structure in which "possible individuals" instantiate "possible properties and relations" in different "possible worlds." I use scare quotes here because these elements of modal models are not to be taken with ontological seriousness: the set of "worlds" in a modal model may contain anything at all—numbers, people, fish, pure sets, and so on. That this will not limit the ersatz-pluriverse account to purely formal uses of worlds will be made clear in section IIIh below.

Leaving out accessibility for simplicity, a *modal structure* will be understood as a 6-tuple $\langle \mathscr{W}, @, \mathscr{D}, \mathscr{P}, \mathscr{Q}, \mathscr{I} \rangle$ where:[24]

- $\mathscr{W}$ is a nonempty set (the set of "worlds").
- $@$ is some member of $\mathscr{W}$ ("the actual world").
- $\mathscr{D}$ is a set (the set of "individuals").
- $\mathscr{P}$ is a set (the set of "properties"). Each member of $\mathscr{P}$ is to have two associated integers. The first integer (≥ 1) is the "number of places" of the property; the second (0 or 1) is the "type" of the property: type 0 for properties of individuals, type 1 for properties of properties.
- No two of $\mathscr{W}$, $\mathscr{D}$, and $\mathscr{P}$ overlap.
- $\mathscr{Q}$ is a function that assigns to each world w an ordered pair $\langle \mathscr{D}_w, \mathscr{P}_w \rangle$, where $\mathscr{D}_w \subseteq \mathscr{D}$ and $\mathscr{P}_w \subseteq \mathscr{P}$. ($\mathscr{D}_w$ and $\mathscr{P}_w$ are to be thought of as the individuals and properties existing in w.) We impose the requirement of world-bound individuals: that if $w \neq w'$ then $\mathscr{D}_w$ and $\mathscr{D}_{w'}$ do not overlap. Overlap of distinct $\mathscr{P}_w$'s is permitted—thus transworld identity of properties is represented.
- $\mathscr{I}$ is a function that assigns to each *n*-place $p \in \mathscr{P}$ an *n*-place intension. If p is type 0, then the intension of p is a function that assigns to each world w a set of *n*-tuples drawn from $\mathscr{D}_w$ (the "extension of p in w"). If p is type 1, then its intension assigns to each w a set

---

[24] In the following, I use unitalicized variables for elements of modal models, to distinguish them from the variables in the possibilist language (and metalinguistic variables for those variables) introduced in section IIId; the latter will be italicized.

of *n*-tuples of members of $\mathcal{P}_w$. The requirement that if p $\notin$ $\mathcal{P}_w$ then the extension of p in w is empty should presumably be imposed.

Both the modal language and the possibilist language may be interpreted in these modal structures. The Lagadonian names are common to both languages, and thus each language may be given the same definition of a *modal model*, which may be understood as a modal structure plus an interpretation function $\mathcal{F}$ that assigns *denotations* to these names:

- To each individual constant, $\mathcal{F}$ assigns some member of $\mathcal{D}_@$.
- To each Lagadonian name for an *n*-place property of type *m* (*m* = 0 or 1), $\mathcal{F}$ assigns some *n*-place member of $\mathcal{P}_@$ of type *m*.

The definition of *truth in* a given modal model must be different for our two languages. For the modal language, the definition is more or less the usual one, subject to the following remarks. As usual, one defines truth-in-a-world recursively; truth in a model is then truth in the actual world of that model. $\ulcorner\alpha_1,\ldots,\alpha_n$ instantiates $\Pi\urcorner$ (relative to an assignment to the variables; this will be suppressed from now on) is true at w if and only if the ordered *n*-tuple of the referents of the terms $\alpha_1,\ldots,\alpha_n$ is in the extension, at w, of the property denoted by Π. (Thus, in a sense, the meaning of the instantiation predicate is "hard-wired" into the definition of truth in a model.) Necessity is truth in all worlds; possibility, truth in some world; actuality, truth in the actual world (@). The language is infinitary, so the usual truth conditions for infinite conjunctions and infinite blocks of quantifiers must be adapted to the modal case in the natural way.[25] The quantifiers are actualist; thus, the individual variables range, at a world w, over $\mathcal{D}_w$, the lower-order property variables over the type 0 members of $\mathcal{P}_w$, and the higher-order property variables over the type 1 members of $\mathcal{P}_w$.[26]

For the possibilist language, truth-in-a-model must be understood differently. Since the possibilist language contains no modal operators, we do not define truth in worlds, but rather truth simpliciter. The modal model is now treated much like an ordinary nonmodal model, with domains for three sorts of variables: the world variables range over $\mathcal{W}$, the individual variables over the whole of $\mathcal{D}$, the property variables over all of the members of $\mathcal{P}$ with the appropriate type. When $\alpha_1,\ldots,\alpha_n$ are terms for individuals, $\ulcorner\alpha_1,\ldots,\alpha_n$ instantiate $\Pi\urcorner$ is true if and only if the denotations of $\alpha_1,\ldots,\alpha_n$ are all in the same world, w, and the n-tuple of these denotations is in the extension in w of the denotation of Π. If $\Pi_1,\ldots,\Pi_n$, and Π are terms for properties, and ω is a term for a world (either a world variable or **@**), then $\ulcorner\Pi_1,\ldots,\Pi_n$ instantiate Π at $\omega\urcorner$ is true if and only if the n-tuple of the denotations of the Π_i's is in the extension at the denotation of ω of the denotation of Π. The predicate 'exists in' and the constant **@** must have "hard-wired" meanings, in the sense that they will always be given the same interpretation in the models:

---

[25] See Dickmann.

[26] We may ignore the fact that the truth definition for sentences with individual quantification across modal operators would ordinarily require counterpart theory. The only modal sentences I shall be examining for truth in modal models are those in CONSTRAINTS, which have no such quantification—see section IIIc.

- The denotation of **@** is ⌐.
- $\ulcorner\alpha$ exists in $\omega\urcorner$, where α is an individual (property) term and ω is a world term, is true if and only if a $\in \mathscr{D}_w$ (if and only if a $\in \mathscr{P}_w$), where a and w are the referents of α and ω, respectively.

We have, then, a single conception of a modal model, but two definitions of truth in a modal model, call them "truth$_p$" and "truth$_m$," for the possibilist and modal languages, respectively. There are two model-theoretic notions of entailment (and thus of equivalence) corresponding to these notions: a sentence in the possibilist (modal) language entails$_p$ (entails$_m$) another if and only if the latter is true$_p$ (true$_m$) in every modal model in which the first is true$_p$ (true$_m$).

The construction has assumed world-bound individuals, to make room for counterpart theory (section IIIf). If this assumption were given up, various changes would need to be made, including adding a place for worlds to the instantiation predicate for properties of individuals, making appropriate revisions to the definition of truth$_m$, and allowing domains of distinct worlds in modal models to overlap.

IIIc. Realistic modal models. Since the ersatz pluriverse is intended to represent the space of *possible* worlds, it had better not imply the existence of worlds with married bachelors, round squares, and the like. Just as linguistic ersatzers make use of a modal notion of consistency, I shall make use of the modal notions of possibility and necessity in constructing the ersatz pluriverse. More specifically, I assume the notion of sentences in the modal language being *true* under a Lagadonian interpretation, or "L-true" for short. On this interpretation, the Lagadonian names (of both individuals and properties and relations) are interpreted as denoting themselves, the instantiation predicate is interpreted as meaning instantiation, and the modal operators $\Box$ and $\Diamond$ are interpreted as meaning necessity and possibility, respectively. L-truth must be sharply distinguished from both truth$_p$ and truth$_m$, each of which holds only relative to modal models. L-truth does not concern modal models, but rather the "real live modal facts."

A certain class, CONSTRAINTS, of L-true sentences in the modal language will be used to construct the ersatz pluriverse. Which sentences, exactly, should be included in CONSTRAINTS? The truth of the members of CONSTRAINTS will be, in essence, "built-into" the ersatz pluriverse. Since I give a counterpart-theoretic account of de re modal sentences below, I do not want truth values for these sentences built into the ersatz pluriverse from the start. Thus, CONSTRAINTS will consist of exactly the L-true de dicto sentences in the modal language—that is, the L-true sentences in the modal language which contain neither (i) individual variables in modal contexts bound by quantifiers outside those contexts, nor (ii) Lagadonian names of actual individuals in modal contexts. (If the theory is not to be coupled with counterpart theory, then CONSTRAINTS may include all L-true sentences of the modal language.)

Call any modal model $\mathcal{M}$ in which every member of CONSTRAINTS is true$_{\mathrm{m}}$ a *realistic* modal model. The modal language allows arbitrarily long infinite formulas, so strictly speaking CONSTRAINTS cannot exist as a set: understand the definition as saying that a realistic model is one in which every sentence in the modal language satisfying the membership condition for CONSTRAINTS is true$_{\mathrm{m}}$. Thus, a realistic modal model is a modal model $\mathcal{M}$ such that for every de dicto sentence, S, in the modal language, if S is L-true then S is true$_{\mathrm{m}}$ *in* $\mathcal{M}$. A realistic model is a model of logical space that is as accurate as possible, given that we have only truths expressible in the modal language to guide us in its construction.

IIId. Pluriverse sentences. Define a *pluriverse sentence* as any maximal description of any realistic model, where a maximal description of $\mathcal{M}$ is a sentence in the possibilist language which is true$_{\mathrm{p}}$ in $\mathcal{M}$ and which entails$_{\mathrm{p}}$ any other sentence in the possibilist language that is true$_{\mathrm{p}}$ in $\mathcal{M}$. More than one sentence will satisfy this definition, so I must cease the pretense of uniqueness.[27]

We need to know that pluriverse sentences do indeed exist. In the remainder of this section, I shall sketch a proof that every realistic model has a maximal description. So assuming that realistic models exist, pluriverse sentences exist.[28] (This section may be skipped with little danger to understanding the rest of this article.)

---

[27] Distinct logically equivalent pluriverse sentences are harmless. It is unclear to me at present whether a harmful multiplicity arises from CONSTRAINTS failing to constrain realistic models up to isomorphism.

We first define the concept of *isomorphic* modal models in the obvious way:

> $\mathscr{M}$ ($= \langle \mathscr{W}, @, \mathscr{D}, \mathscr{P}, \mathscr{Q}, \mathscr{I} \rangle$ + interpretation $\mathscr{F}$) is isomorphic to $\mathscr{M}'$ ($= \langle \mathscr{W}', @', \mathscr{D}', \mathscr{P}', \mathscr{Q}', \mathscr{I}' \rangle$ + interpretation $\mathscr{F}'$) if and only if there exists a one-one mapping, f, such that:
>
> (i) f maps $\mathscr{W}$ onto $\mathscr{W}'$, $\mathscr{D}$ onto $\mathscr{D}'$, and $\mathscr{P}$ onto $\mathscr{P}'$
> (ii) $f(@) = @'$
> (iii) $i \in \mathscr{D}_w$ if and only if $f(i) \in \mathscr{D}'_{f(w)}$, for any $i \in \mathscr{D}$ and any $w \in \mathscr{W}$
> (iv) $i \in \mathscr{P}_w$ if and only if $f(i) \in \mathscr{P}'_{f(w)}$, for any $i \in \mathscr{P}$ and any $w \in \mathscr{W}$
> (v) $\langle i_1,\dots,i_n \rangle \in \mathscr{I}(j)(w)$ if and only if $\langle f(i_1),\dots,f(i_n) \rangle \in \mathscr{I}'(f(j))(f(w))$, for any $w \in \mathscr{W}$, any n-place $j \in \mathscr{P}$ of type 0, and any $i_1,\dots,i_n$ in $\mathscr{D}_w$
> (vi) $\langle i_1,\dots,i_n \rangle \in \mathscr{I}(j)(w)$ if and only if $\langle f(i_1),\dots, f(i_n) \rangle \in \mathscr{I}'(f(j))(f(w))$, for any $w \in \mathscr{W}$, any n-place $j \in \mathscr{P}$ of type 1, and any $i_1,\dots,i_n$ in $\mathscr{P}_w$
> (vii) $\mathscr{F}'(\beta) = f(\mathscr{F}(\beta))$, for every property or individual constant β

The following will be proved below:

> Theorem: for any realistic modal model $\mathscr{M}$, there exists a sentence ϕ of the possibilist language (the "canonical pluriverse sentence for $\mathscr{M}$") which is true$_p$ in $\mathscr{M}$, and which is such that if it is true$_p$ in an arbitrary modal model $\mathscr{M}'$ then $\mathscr{M}'$ is isomorphic to $\mathscr{M}$.

It is obvious that the same sentences in the possibilist language will be true$_p$ in isomorphic models. Hence, it follows from the theorem that in any modal model there is a true$_p$ sentence, in "canonical form," that entails$_p$ every other sentence in the possibilist language that is true$_p$ in that model—a pluriverse sentence for that model. Given the definition of a pluriverse sentence, it then follows immediately that any pluriverse sentence is equivalent$_p$ to some pluriverse sentence in canonical form.

To prove the theorem, consider any modal model $\mathscr{M}$ ($= \langle \mathscr{W}, @, \mathscr{D}, \mathscr{P}, \mathscr{Q}, \mathscr{I} \rangle$ + interpretation $\mathscr{F}$). The following notation will be used:

> $\exists(w_i{:}\dots w_i\dots)$: the (possibly infinitary) existential quantification of the class of variables w_i satisfying condition...w_i.... Similarly for variables of other types, and for universal quantification.

---

[28] One challenge to the existence of realistic models is discussed in section IVa. A separate challenge: realistic models will not exist, given the current definitions, if the logic of the sentences in CONSTRAINTS is not S5. In that case, an accessibility relation would need to be introduced into realistic models, a predicate for accessibility added to the possibilist language, and the definitions of truth$_m$ and truth$_p$ adjusted accordingly.

$\bigwedge(\phi\text{:}...\phi...)$: the (possibly infinitary) conjunction of the class of formulas ϕ satisfying condition...ϕ.... Similarly for disjunction: $\bigvee(\phi...\phi...)$. Let the disjunction of the empty set be some logically false sentence, such as $\exists x\, x \neq x$; let the conjunction of the empty set be some logically true sentence.

w, w_i, and so on, are world variables; x, y, x_i, and so on, are individual variables; P, P_i, Q, and so on, are property variables (of either higher or lower type).

Every member of $\mathscr{D}_{\nearrow}$ is denoted by (that is, is assigned by $\mathscr{F}$ to) some individual constant: since we have an individual Lagadonian constant for *every* (actual) individual, the following sentence in the modal language is L-true, and so must be true$_m$ in $\mathscr{M}$, and so must be true$_m$ at $\nearrow$:

$$\forall x \bigvee(x = \alpha\text{:}\ a \text{ is an individual constant})$$

Since the constants are Lagadonian, the sentence $\ulcorner a \neq b \urcorner$ is L-true and so true$_m$ in $\mathscr{M}$ whenever a and b are distinct individual constants. Thus, every member of $\mathscr{D}_{\nearrow}$ is denoted in $\mathscr{M}$ by a unique individual constant. It can be shown similarly that every member of $\mathscr{P}_{\nearrow}$ is denoted by a unique property constant. Next:

For each i $\in \mathscr{W}$ other than $\nearrow$, introduce a distinct world variable w_i.
For each i $\in \mathscr{D}$ but not in $\mathscr{D}_{\nearrow}$, introduce a distinct individual variable x_i.
For each i $\in \mathscr{P}$ but not in $\mathscr{P}_{\nearrow}$, introduce a distinct property variable P_i.
For each i $\in \mathscr{D}_{\nearrow}$, call the individual constant denoting *i* in $\mathscr{M}$ "a_i."
For each i $\in \mathscr{P}_{\nearrow}$, call the individual constant denoting *i* in $\mathscr{M}$ "p_i."

To simplify the various special cases that arise in connection with the actual world $\nearrow$, let us introduce the following abbreviations:

For any i $\in \mathscr{P}$, let "Π_i" denote either the constant p_i or the variable P_i, depending on whether i is, or is not, in $\mathscr{P}_{\nearrow}$.
For any i $\in \mathscr{D}$, let "α_i" denote either the constant a_i or the variable x_i, depending on whether i is, or is not, in $\mathscr{D}_{\nearrow}$.
For any i $\in \mathscr{W}$, let "ω_i" denote either the constant @ or the variable w_i, depending on whether i is, or is not, $\nearrow$.

We now construct the required pluriverse sentence ϕ as follows:

$$\phi = \ulcorner \exists(w_i\text{: i} \in \mathscr{W} \text{ but i} \neq \nearrow)\exists(x_i\text{: i} \in \mathscr{D} \text{ but i} \notin \mathscr{D}_{\nearrow})\exists(P_i\text{: i} \in \mathscr{P} \text{ but i} \notin \mathscr{P}_{\nearrow})\ [\text{DISTINCTNESS \& COMPLETENESS \&} \bigwedge(W_i\text{: i} \in \mathscr{W})]\urcorner$$

where

$$\text{DISTINCTNESS} = \ulcorner \bigwedge(\omega_i \neq \omega_j\text{: i, j} \in \mathscr{W} \text{ and i} \neq \text{j}) \ \&\ \bigwedge(\alpha_i \neq \alpha_j\text{: i, j} \in \mathscr{D} \ \&\ \text{i} \neq \text{j}) \ \&\ \bigwedge(\Pi_i \neq \Pi_j\text{: i, j} \in \mathscr{P} \ \&\ \text{i} \neq \text{j})\urcorner$$

and

$$\text{COMPLETENESS} = \ulcorner \forall w \bigvee(w = \omega_i : \mathrm{i} \in \mathscr{W}) \ \&\ \forall x \bigvee(x = \alpha_i : \mathrm{i} \in \mathscr{D}) \ \&\ \forall P \bigvee(P = \Pi_i : \mathrm{i} \in \mathscr{P})\urcorner$$

and

$$W_i = \ulcorner \forall x[x \text{ exists in } \omega_i \leftrightarrow \bigvee(x = \alpha_j : \mathrm{j} \in \mathscr{D}_i)] \ \&\ \forall P[P \text{ exists in } \omega_i \leftrightarrow \bigvee(P = \Pi_j : \mathrm{j} \in \mathscr{P}_i)]$$

$$\&\ \bigwedge[\forall y_1,\ldots,\forall y_n((y_1 \text{ exists in } \omega_i \ \&\ldots\&\ y_n \text{ exists in } \omega_i \ \&\ y_1,\ldots, y_n \text{ instantiate } \Pi_j) \leftrightarrow \bigvee(y_1 = \alpha_{j1} \ \&\ldots\&\ y_n = \alpha_{jn} : \langle \mathrm{j}_1,\ldots,\mathrm{j}_n\rangle \in \mathscr{I}(\mathrm{j})(\mathrm{i})) : \mathrm{j} \text{ is an } n\text{-place type 0 member of } \mathscr{P}]$$

$$\&\ \bigwedge[\forall Q_1,\ldots,\forall Q_n((Q_1,\ldots,Q_n \text{ instantiate } \Pi_j \text{ at } \omega_i \leftrightarrow \bigvee(Q_1 = \Pi_{j1} \ \&\ldots\&\ Q_n = \Pi_{jn} : \langle \mathrm{j}_1,\ldots,\mathrm{j}_n\rangle \in \mathscr{I}(\mathrm{j})(\mathrm{i})) : \mathrm{j} \text{ is an } n\text{-place type 1 member of } \mathscr{P}]\urcorner$$

Remarks: DISTINCTNESS ensures that no two property or individual terms, whether variables or constants, denote the same thing. COMPLETENESS says that there are no worlds other than those denoted by $@$ and the variables w_i, and no individuals or properties other than those denoted by the Lagadonian constants and the variables x_i and P_i. The first component of the world conjunct W_i says that the individuals and properties in world i are exactly those denoted by the terms α_j and Π_j. The second component contains an infinite conjunction, one conjunct for each lower-order property; each of these conjuncts says, for the property in question, that the individuals that satisfy the property in i are exactly such and such. The third does the same thing for all of the higher-order properties.

It is tedious but straightforward to verify that any model in which ϕ is true$_p$ is isomorphic to $\mathscr{M}$.

IIIe. The proposal: linguistic version. Let S be any sentence of the possibilist language, and think of S under the Lagadonian interpretation, as a sentence concerning reality, not any modal model. S might be the sort of sentence that proves so useful in philosophy or linguistics, quantifying over nonactual worlds, individuals, and properties. Only a modal realist like Lewis could admit the truth of such a sentence—hence the need for a reduction. So:

> Proposal (linguistic version): reinterpret S as the assertion that S is entailed$_p$ by all pluriverse sentences.

Since the proposal generates truth conditions only for sentences stated within the Lagadonian possibilist language, English sentences about possibilia must be regimented in that language. For example, an English subject-predicate atomic sentence $\ulcorner Fa \urcorner$ may be translated as the Lagadonian sentence $\ulcorner \alpha$ instantiates $\pi \urcorner$, where α is a Lagadonian name for the referent of a, and π is a Lagadonian name for the property expressed by F.

Let us examine pluriverse sentences in more depth.

First, pluriverse sentences need not have the logical form introduced in section II:

> THERE ARE worlds w_1, w_2,...and THERE ARE properties and relations P_1, P_2,...that are distinct from the following actual properties and relations:..., and THERE ARE possible individuals x_1, x_2,...that are distinct from the following actual individuals:..., SUCH THAT:...w_1...and...w_2...and....

beginning with quantifiers for the worlds, properties, and individuals, and concluding with an infinite conjunction of open sentences—world conjuncts— one for each world variable. Maximal descriptions of modal models need not have this logical form. As I argued in the previous section, however, every pluriverse sentence is equivalent$_p$ to some such sentence, which I call a *canonical pluriverse sentence*. It is harmless, therefore, to think of pluriverse sentences as having this canonical form.

Second, no pluriverse sentence can assert the existence of a world in which impossibilities occur. Suppose a world conjunct in some pluriverse sentence contained 'x exists at w' and 'x is a married bachelor'. Any pluriverse sentence is true$_p$ in some realistic model; so some individual in the domain of some world of some realistic model would be in the extension of 'married bachelor' in that world. But this is impossible, since the sentence '$\Box$ there exist no married bachelors' is L-true, and so is a member of CONSTRAINTS, and so must be true$_m$ in any realistic model.

Third, every possibility will be represented as holding in some world, by any pluriverse sentence. Where $\mathcal{M}$ is any realistic model, if $\ulcorner \Diamond\phi \urcorner$ is an L-truth of the modal language then it must be true$_m$ in $\mathcal{M}$, in which case ϕ will be true$_m$ in some world of $\mathcal{M}$. But then any pluriverse sentence based on $\mathcal{M}$ will entail$_p$ a sentence asserting that ϕ holds in some world. Notice that ϕ might be an infinitary sentence since the modal and possibilist languages allow such sentences; thus, there is no restriction to finitely stateable possibilities. Relatedly, pluriverse sentences will represent necessary truths as holding in every possible world. Thus, on the pluriverse view, truths about possibility and necessity in the modal language "mesh" with their worlds translations.

Fourth, the possible worlds represented by the world conjuncts in pluriverse sentences are "completely specific." Suppose a pluriverse sentence ϕ entails$_p$ the following:

> There exists a world, an individual x existing at that world, and properties $P_1,\ldots,P_n$ and Q existing at that world, such that x has properties $P_1,\ldots,P_n$.

It must then entail$_{\text{p}}$ one of the following:

> There exists a world, an individual x existing at that world, and properties $P_1,\ldots,P_n$ and Q existing at that world, such that x has properties $P_1,\ldots,P_n$ as well as Q.
>
> There exists a world, an individual x existing at that world, and properties $P_1,\ldots,P_n$ and Q existing at that world, such that x has properties $P_1,\ldots,P_n$, but not Q.

For if ϕ is a pluriverse sentence, then it is a maximal description of some modal model $\mathcal{M}$; if ϕ entails$_{\text{p}}$ the first sentence, then that sentence is true$_{\text{p}}$ in $\mathcal{M}$; but then either the second or third sentence must be true$_{\text{p}}$ in $\mathcal{M}$, and hence one must be entailed$_{\text{p}}$ by ϕ.[29]

Fifth, as constructed in the previous section, each pluriverse sentence in canonical form contains a clause COMPLETENESS asserting that its inventory of worlds, individuals, and properties is complete:

> THERE ARE worlds w_1, w_2,...and THERE ARE properties and relations P_1, P_2,...that are distinct from the following actual properties and relations:..., and THERE ARE possible individuals x_1, x_2,...that are distinct from the following actual individuals:..., SUCH THAT: COMPLETENESS and...w_1...and...w_2...and....

The clause COMPLETENESS looks like this:

> COMPLETENESS: there are no worlds other than @, w_1, w_2,..., and there are no properties or relations other than charge, charm, [list all actual properties and relations], and P_1, P_2,..., and there are no possible individuals other than Socrates, Plato, and Aristotle, [list all actual individuals], and x_1, x_2,....

If a pluriverse sentence entailed$_{\text{p}}$ no such clause it could not be a *maximal* description of any realistic model, for it would fail to entail$_{\text{p}}$ many of the universally quantified sentences true$_{\text{p}}$ in that realistic model.

Sixth, each pluriverse sentence will include information about the actual world. For any *n*-place property Π and any n objects $\alpha_1,\ldots,\alpha_n$ that, in fact, instantiate Π, the atomic sentence in the modal language, $\ulcorner\alpha_1,\ldots,\alpha_n$ instantiate $\Pi\urcorner$, is L-true, and so is in CONSTRAINTS, and so is true$_{\text{m}}$ in any realistic modal model. Let $\mathcal{M}$ be any realistic model, and let @ be the actual world of $\mathcal{M}$. Given the definition of a modal model, $\mathcal{D}_@$ contains the denotations $a_1,\ldots,a_n$ of $\alpha_1,\ldots,\alpha_n$, and $\mathcal{P}_@$ contains p, the denotation of Π. By the definition of truth$_m$, $\langle a_1,\ldots, a_n\rangle$ is in the extension of p at @ in $\mathcal{M}$. Thus, in any realistic modal model, the actual world @ will encode all the nonmodal truths about the (real) actual world. This information finds its way into pluriverse sentences as follows. Given what we have said, the following sentence of the possibilist language must be true$_p$ in any realistic model:

---

[29] I ignore vagueness throughout.

(@) Π exists in @, $\alpha_1,\ldots,\alpha_n$ exist in @, and $\alpha_1,\ldots,\alpha_n$ instantiate Π.

It must, therefore, be entailed$_p$ by any pluriverse sentence. Therefore: L-true nonmodal subject-predicate propositions are reported by pluriverse sentences as holding in the actual world. (Each pluriverse sentence ϕ in canonical form will collect all the information contained in sentences like (@) into a single *actual world conjunct*—the one of the world conjuncts that completely describes @, the actual world of the model on which ϕ is based.)

Seventh, pluriverse sentences say more about worlds than they say about them individually; they do not merely entail$_p$ sentences of the form ⌜There exists a world such that ϕ⌝. Suppose that property variable P occurs in world conjuncts W_1 and W_2 (which, in turn, correspond to world variables w_1 and w_2) in some pluriverse sentence. It will then entail$_p$:

There exists a property P, and worlds w_1 and w_2, such that W_1 and W_2.

This sort of consequence, a claim about two worlds "from an external perspective," is what makes possible the solution to the problem of descriptive power.

IIIf. Inter-world facts. Though the bare-bones pluriverse theory is now in place, several additions are needed to yield a complete theory of worlds. The present section will sketch, in considerably less detail than the preceding sections, how some of this might be carried out.

To gain the full benefits of talk of possibilia, "paradise on the cheap" as Lewis says, we want to speak, not only of possibilia, but also of set-theoretic constructions of possibilia. Chalmers's two-dimensional framework, for example, requires functions defined over possible worlds and individuals. Ersatz pluriverses must therefore represent *sets* of worlds, individuals, and properties in addition to the worlds, individuals, and properties themselves. To accomplish this, the language of set theory could be added to the possibilist language, and the definition of truth$_p$ in a modal model modified so that sentences containing set-theoretic vocabulary are true$_p$ in $\mathcal{M}$ if and only if they accurately describe $\mathcal{M}$'s set-theoretic structure, in the following sense. $\mathcal{M}$'s set of "worlds," $\mathcal{W}$, is a certain set, and thus has a certain set-theoretic structure: it has certain subsets, which themselves have subsets, and so on. Similarly, $\mathcal{M}$'s set of individuals, $\mathcal{D}$, has a certain set-theoretic structure. $\mathcal{W}$ and $\mathcal{D}$ do not need to be augmented to represent their set-theoretic structures; they simply *have* their structures. The definition of truth$_{\text{p}}$ must be modified so that a set-theoretic sentence in the possibilist language turns out true$_{\text{p}}$ in $\mathcal{M}$ if and only if it correctly represents the set-theoretic structure that $\mathcal{M}$ *has*. Thus, the very same pluriverse sentences defined above will now entail$_{\text{p}}$ set-theoretic sentences, despite containing no set-theoretic vocabulary. Sentences about sets of possibilia will be entailed$_{\text{p}}$ by any pluriverse sentence, and so will turn out true, on the pluriverse view.

Facts about sets of possibilia are "inter-world" facts—facts not from the perspective of any one world, but rather from an external perspective. There are other examples of such facts: (i) the pluriverse should, perhaps, represent the existence of transworld mereological sums; (ii) the pluriverse should, perhaps, represent relations holding between inhabitants of worlds and transworld sets or sums: for example, belief relations between persons and transworld sets thought of as propositions; (iii) the pluriverse should represent the holding of a counterpart relation between objects from different worlds. To achieve these goals, the possibilist language must be enriched, modal models must be augmented to represent the desired facts, and the definition of truth$_{\text{p}}$ must be adjusted accordingly.

The addition of counterpart theory to the pluriverse view will illustrate some of the complications that can arise with these additions. Where $\mathcal{M}$ is any modal model and $\mathcal{C}$ is a binary relation over $\mathcal{M}$'s domain, $\mathcal{D}$, call $\langle \mathcal{M}, \mathcal{C} \rangle$ a *C-model*. $\mathcal{C}$ is to be thought of as the counterpart relation for $\mathcal{M}$. (A refinement would be to allow multiple counterpart relations.[30])

We now need the notion of a *realistic C-model*. Since the counterpart relation is a similarity relation (though not necessarily a relation of intrinsic similarity), whether it holds depends on what properties are instantiated by its relata and what relations its relata bear to their worldmates. Thus, for $\langle \mathcal{M}, \mathcal{C} \rangle$ to be realistic, it is not enough for $\mathcal{M}$ to be realistic; in addition, the holding of $\mathcal{C}$ must "mesh" with the properties and relations instantiated by the objects in that model. This is not specific to counterpart theory: *inter*-world relations will usually be constrained by the *intra*-world facts about their relata.

---

[30] See Lewis, "Counterparts of Persons and Their Bodies," this JOURNAL, LXVIII, 7 (April 8, 1971): 203-11.

In fact, the counterpart relation seems *determined* by intra-world facts. (Whether this is true in general, or whether some interesting inter-world relations are constrained without being determined by intra-world facts, is debatable.) Thus, there should be a function, *c-determination*, which when applied to a realistic modal model $\mathcal{M}$ yields the appropriate counterpart relation, $\mathscr{C}$, over the individuals of $\mathcal{M}$. Given this function, we can then define the notion of a realistic C-model: $\langle \mathcal{M}, \mathscr{C} \rangle$ is realistic if and only if $\mathcal{M}$ is realistic and $\mathscr{C} =$ c-determination($\mathcal{M}$). Thus, our problem of defining 'realistic C-model' reduces to the problem of defining 'c-determination'.

The latter would be easy if we had a particular *theory* of the counterpart relation. Suppose, for example, that x is a counterpart of y if and only if x and y both instantiate a certain property P. Then we could define c-determination($\mathcal{M}$) as the relation that holds between members of $\mathcal{M}$'s domain x and y if and only if both x and y are in the extension of P in their respective worlds. But the counterpart relation is often taken as an undefined primitive (although explained intuitively in terms of similarity).

One *could* replace the primitive counterpart relation with a new primitive, that of c-determination. C-determination could be explained informally in terms of similarity and so would count as a kind of similarity primitive. But C-determination would be an unwieldy primitive. It would be better to define it in terms of (the relevant sort of) similarity itself, and then take the latter as primitive instead. Under certain assumptions this is, in fact, possible.

Here is a sketch of a definition of 'c-determination' in terms of a similarity predicate 'c-similar'. Consider any modal model, $\mathcal{M}$, containing individuals x_1 and x_2, from worlds w_1 and w_2, respectively. Form a pair of open formulas W_1 and W_2 of the possibilist language describing w_1 and w_2 as follows. Represent each of the individuals and properties in the two worlds with distinct terms, letting any properties from the actual world of the model be represented by the Lagadonian constants that denote those properties in the model, and letting all individuals (whether in the actual world of the model or not) and all nonactual properties be represented by variables. Let x_1 and x_2 be the variables corresponding to individuals x_1 and x_2, respectively; let $\alpha_1,\ldots,$be a list of all the variables other than x_1 used to represent things from w_1, and let $\beta_1,\ldots,$be a list of all variables other than x_2 used for w_2. Next, in the manner of section IIId, using these variables construct (possibly infinite) open formulas W_1 and W_2 that completely describe the pattern of instantiation of properties within worlds w_1 and w_2. We now say that c-determination($\mathcal{M}$) holds between $\mathbf{x}_1$ and $\mathbf{x}_2$ if and only if the following is an L-true sentence of the modal language:

> Necessarily, for all x_1, α_1,..., and for all distinct x_2, β_1,..., if W_1 and W_2, then x_1 and x_2 are c-similar.

("any objects that are as W_1 and W_2 say that x_1 and x_2 are would necessarily be c-similar"). A potential problem, however, is that W_1 and W_2 may not be satisfiable by distinct sets of objects—perhaps there is no one possible world containing wholly distinct parts that are as described by W_1 and W_2. In this case, the definition will hold vacuously, in virtue of the antecedent of the conditional, ⌜if W_1 and W_2, then...⌝, being necessarily false. W_1 and W_2 express massive conditions that are satisfied by all the objects in some possible world of some modal model.[31] The following principle would ensure that the definition never holds vacuously: for any two possible worlds, there exists another possible world containing distinct duplicates of the first two worlds. If the principle is true, then the conditionals in the definition will not be vacuous; for even though W_1 and W_2 describe two entire possible worlds, there will *also* be a third world containing two parts, one with objects that satisfy W_1, the other with objects that satisfy W_2. But this principle is controversial.

Reliance on this principle could be avoided if we were willing to accept nonvacuous counterfactual conditionals with metaphysically impossible antecedents[32]: we could then say that c-determination($\mathscr{M}$) holds between $\mathbf{x}_1$ and $\mathbf{x}_2$ if and only if the following is L-true:

> If there had existed x_1, α_1,..., and distinct x_2, β_1,..., such that W_1 and W_2, then it would have been the case that x_1 and x_2 are (relevantly) similar.

The cost here would be accepting this counterfactual connective as a new modal primitive.

Whether 'c-determination' is a primitive or is defined in terms of 'c-similar', realistic models can represent the holding of a counterpart relation over possibilia.

IIIg. The proposal: nonlinguistic version. Realistic models were used to define pluriverse sentences, which then were used to reinterpret sentences about possibilia. But one could bypass the pluriverse sentences and utilize realistic models directly:[33]

---

[31] Compare Lewis, *On the Plurality of Worlds*, p. 238.

[32] I think we do make nonvacuous sense of at least some counterpossibles, though it is an open question whether there are counterpossibles rich enough to ground the proposed definition. See Jeffrey Goodman, "Extended Ersatz Realism," forthcoming in *Pacific Philosophical Quarterly*; William Lycan, *Modality and Meaning* (Boston: Kluwer, 1994), pp. 38-39; and Takashi Yagisawa, "Beyond Possible Worlds," *Philosophical Studies*, LIII (1987): 175-204.

> Proposal (nonlinguistic version): reinterpret S as the claim that S is true$_p$ in every realistic model.

Given the theorem of section IIId, the linguistic and nonlinguistic versions of the proposal are equivalent.[34] The nonlinguistic version is simpler, but I find the linguistic version more intuitive. The linguistic version also facilitates comparison with other theories: for instance, linguistic ersatzism and modal fictionalism (to be discussed below).

IIIh. Taking stock. Our inquiry into the ontology of worlds has been driven by the dilemma: Lewisian modal realism is unacceptable, yet possibilia are useful in semantics and ontology. My solution is the proposal sketched in the preceding sections. The proposal is superficially similar to the use of Kripke models as "mere structures" since the particular entities in the sets of "worlds," "individuals," and "properties" in modal models are insignificant; these sets may contain, for example, numbers, persons, or fish. But as mentioned at the outset, the purely formal use of Kripke models is severely limited. More interesting uses for possible worlds require the ability to speak of particular worlds, classes of worlds, functions defined on worlds, and so on. The proposal I have given allows this, by providing a reduction of sentences like:

---

[33] This formulation is due to the extremely valuable suggestions of Kit Fine, to whom I am greatly indebted in my search for an adequate formal construction of the ersatz pluriverse.

[34] To prove: S is true$_p$ in all realistic models if and only if S is entailed$_p$ by all pluriverse sentences. First consider the proposal before the introduction of set-theoretic vocabulary discussed in IIIf.

Let S be true$_p$ in all realistic models, let PS be a pluriverse sentence, and suppose that PS is true$_p$ in some modal model $\mathscr{M}$. Since PS is a pluriverse sentence, it is a maximal description of some realistic model $\mathscr{M}'$. S is true$_p$ in $\mathscr{M}'$, and so is entailed$_p$ by PS, and so is true$_p$ in $\mathscr{M}$.

On the other hand, suppose S is entailed$_p$ by all pluriverse sentences, and let $\mathscr{M}$ be a realistic model. By the theorem of IIId, some pluriverse sentence is true$_p$ in $\mathscr{M}$; it entails$_p$ S, and so S is true$_p$ in $\mathscr{M}$ as well.

Now consider the introduction of set-theoretic vocabulary from IIIf. The proof just given relies only on (i) definitions which have not changed, and (ii) the theorem of IIId, that in each realistic modal model there is a true$_p$ pluriverse sentence. Thus, all we must show is that the theorem holds under the new definition of truth$_p$. That theorem established the existence of a sentence ϕ that characterizes a given modal model up to isomorphism. But isomorphic models have the same set-theoretic structure, and hence make the same sentences of the new, set-theoretic possibilist language true$_p$. Thus, ϕ remains a pluriverse sentence for the model, under the new definition.

Notice that ϕ now has new entailments$_p$ (namely, set-theoretic sentences) without having anything added to it, given the new definition of entailment$_p$. Pluriverse sentences could not have characterized all of the set-theoretic facts by name, via Lagadonian names of all the sets in the hierarchy generated by a given model, for then pluriverse sentences could not exist.

Pluriverse sentences still do not entail$_p$ statements about sets existing *in* worlds. How then can the L-truth of '$\Box$(the empty set exists)' be accommodated? Probably the best solution is to follow Lewis in regarding the worlds translation of such a sentence as involving a quantifier over sets that exist "from the point of a world" rather than *in* the world; see postscript A to "Counterpart Theory and Quantified Modal Logic," in Lewis's *Philosophical Papers, Volume 1* (New York: Oxford, 1983), pp. 39-40.

- There is a world in which a donkey talks.
- There exists a pair of worlds and a pair of nonactual fundamental properties such that the worlds differ only by the properties swapping nomic roles.
- No worlds that lack alien fundamental properties differ without differing physically.
- There exists a primary intension of the term 'water'—a function that assigns to any world, *w*, the set of things that would count as water in *w* when *w* is "considered as actual."

The reduction does indeed assign truth conditions to these sentences via modal models whose classes of "worlds," "individuals," and "properties" may contain numbers, persons, fish, and so on; but the important thing is that the sentences are indeed assigned truth conditions. Thus, my reduction allows one to partake fully of modal metaphysics and semantics.

The construction of realistic modal models assumes the notion of truth (under the Lagadonian interpretation) for de dicto sentences in the modal language, and hence assumes de dicto necessity and possibility (which are, of course, interdefineable). These modal notions remain primitive here. Thus, my account of worlds cannot be employed in a reductive account of these modal notions themselves. As explained in section I, a reduction of talk of possibilia which employs primitive possibility and necessity is nevertheless valuable since talk of possibilia runs beyond what can be said in the language of quantified modal logic. The pluriverse account thus reduces talk about possibilia to de dicto sentences in the modal language.

Moreover, as explained in section IIIf, given a similarity primitive (whether c-determination or a similarity relation over individuals), the holding of a counterpart relation can be built into realistic models. The pluriverse account thus yields a theory of the truth conditions for statements about a counterpart relation holding over possibilia. Given the counterpart-theoretic account of de re modality in terms of such statements, truth conditions for de re modal sentences can be given. The usual counterpart-theoretic reduction of de re modality to de dicto modality plus similarity is thereby achieved.

IV PROBLEMS FOR THE PLURIVERSE VIEW

IVa. Cardinality problems.[35] The pluriverse view runs into trouble if there are too many possible worlds or individuals to form a set. Suppose, for example, that there is no upper bound, not even an infinite upper bound, to how many (concrete) individuals may possibly exist. Then there can be no realistic models, for any modal model $\mathcal{M}$ has a domain of individuals, $\mathcal{D}$, which is a set with some cardinality; but where κ is some larger cardinal, there will be an L-true sentence in the modal language saying that it is possible that there are least κ objects[36]; such a sentence would be in CONSTRAINTS, but would not be true$_m$ in $\mathcal{M}$. Any modal model, no matter how large, would misrepresent the pluriverse as containing an upper bound on the size of worlds. A similar problem would arise if there were no upper bound to the number of properties that could exist.

The very first point to make about the cardinality problem is that it is not particular to the pluriverse view. *Every* theory of worlds encounters trouble in this area. The linguistic ersatzist, for example, may admit arbitrarily large worlds, but cannot admit a world with so many individuals that they cannot all be members of a set (except in special cases where the objects display symmetries allowing simpler description); for a linguistic-ersatz world is a maximal consistent *set* of sentences, and sentences themselves are also sets. That is a limitation on possibility, although a bit less severe than an upper bound on world size.[37] Similar problems confront other views that identify possible worlds with abstract entities other than sets of sentences. This is sometimes less easy to see than with linguistic ersatzism, since defenders of these views do not always supply as rigorous assumptions about the behavior of these abstract entities as do the axioms of set theory, but the problem is no less real for that.[38]

---

[35] I thank Daniel Nolan for helpful comments on this section.

[36] Given the notation of section IIId, this sentence could be the following, where S is a set of κ many variables: $\exists(x: x \in S) \wedge (x \neq y: x \in S, y \in S, x$ and y are distinct variables).

[37] Moreover, some plausible-looking recombination principles would generate proper-class-sized worlds from the claim that there are proper-class many possible individuals. One presupposes "transworld identity": for any possible individuals (perhaps drawn from different possible worlds), there is a possible world containing all of those individuals. Another drops that presupposition: for any possible individuals (again perhaps from different worlds), there is a possible world containing distinct duplicates of those individuals.

[38] See Chihara, pp. 120-41.

Moreover, anyone (even a Lewisian modal realist) who believes in possible worlds and individuals as entities, faces the problem of reconstructing applications of possible-worlds talk if there are too many possible worlds or individuals to form a set. If there is no upper bound on the size of possible worlds, then there can be no set of all the possible individuals and no set of all the possible worlds (ersatz or genuine). Therefore, applications of worlds talk (for example, defining semantic values) which require functions defined on the space of possible worlds and individuals are in trouble. Everyone faces the cardinality problem. Below I sketch a few potential lines of response; but the question is complex and calls for further study.[39]

A flat-footed response would be to simply deny (as Lewis did at one point) that there can be arbitrarily many individuals.[40] Although theoretically simple, this response requires an unattractive restriction on what is possible. A theme of the present discussion has been that the theory of worlds should not put controversial constraints on what is possible.

Alternatively, one could try "technical tricks" of various sorts, in various combinations. A drastic move would be to invoke a nonstandard logic or set theory on which Cantorian paradoxes with the universal set do not arise.[41] Less drastically, one might invoke class theory,[42] and allow $\mathscr{W}$, $\mathscr{D}$, and $\mathscr{P}$, in modal models to be proper classes. Since proper classes are usually not allowed to be members of further classes, some further dancing will be needed. For example, if $\mathscr{W}$, $\mathscr{D}$, and $\mathscr{P}$ are proper classes then they cannot be members of a modal structure $\langle \mathscr{W}, \nearrow, \mathscr{D}, \mathscr{P}, \mathscr{Q}, \mathscr{S} \rangle$. One could give up on modal structures as entities, and cash out talk of modal models as *plural* talk of $\mathscr{W}$, $\nearrow$, $\mathscr{D}$, $\mathscr{P}$, $\mathscr{Q}$, and $\mathscr{S}$, plus an interpretation $\mathscr{F}$. While this would not rescue pluriverse sentences, it would at least allow for a redefinition of "true$_{\mathrm{P}}$ in all modal models" and thus would rescue the nonlinguistic version of the theory. If $\mathscr{W}$, $\mathscr{D}$, and $\mathscr{P}$ are proper classes, however, then the existence of $\mathscr{Q}$ and $\mathscr{S}$ is in jeopardy since they are functions defined on $\mathscr{W}$, $\mathscr{D}$, and $\mathscr{P}$. There is no problem if the domain of a world can never be proper-class-sized, since then there would be no need for the values of $\mathscr{Q}$ and $\mathscr{S}$ to be classes; but if in addition to arbitrarily large set-sized worlds we wish to admit (as we probably should) proper-class-sized worlds, then $\mathscr{Q}$ and $\mathscr{S}$ cannot be functions construed as classes of pairs. We might now pull mereology from the bag of tricks. $\mathscr{Q}$ originally was a function that assigns to any $w \in \mathscr{W}$ a pair $\langle \mathscr{D}_w, \mathscr{P}_w \rangle$ of subsets of $\mathscr{D}$ and $\mathscr{P}$. Suppose we accept a mereology of classes on which the following is coherent to require of realistic models: the members of $\mathscr{W}$, the subclasses of $\mathscr{D}$ and the subclasses of $\mathscr{P}$ never overlap mereologically. We could then define $\mathscr{Q}$ as a class, each member of which is the mereological sum of a member of $\mathscr{W}$, a subclass of $\mathscr{D}$, and a subclass of $\mathscr{P}$. For a given $w \in \mathscr{W}$, $\mathscr{D}_w$ could then be recovered—it would be the subclass of $\mathscr{D}$ such that it and w are both part of X, for some $X \in \mathscr{Q}$; similarly for $\mathscr{P}_w$. The case of $\mathscr{S}$ is more complicated. Whether the entire pluriverse theory can be reconstructed along these lines is an open question.

---

[39] See Nolan (chapter 6) for more work on this topic. Does the modal fictionalist face the cardinality problem? Answering this requires deciding whether the fiction must be logically coherent; if not then it could include the claim that there exists a set of all possible worlds and individuals, despite the resulting contradictions.

[40] *On the Plurality of Worlds*, pp. 102-04. For contrary arguments, see Bricker, "Plenitude of Possible Structures," this JOURNAL, LXXXVIII, 11 (November 1991): 607-19; and section IV of Nolan, "Recombination Unbound," *Philosophical Studies*, LXXXIV (1996): 239-62.

[41] See, for example, Quine, "New Foundations for Mathematical Logic," in his *From a Logical Point of View*, pp. 80-101; and Greg Restall, "A Note on Naive Set Theory in LP," *Notre Dame Journal of Formal Logic*, XXXIII (1992): 422-32.

[42] See Paul Bernays, "A System of Axiomatic Set Theory," in G. Muller, ed., *Sets and Classes* (Amsterdam: North-Holland, 1976), pp. 1-119; and Kurt Gödel, *The Consistency of the Axiom of Choice and of the Generalized Continuum-Hypothesis with the Axioms of Set Theory* (Princeton: University Press, 1940).

Yet another response leaves the definition of a modal model intact, but alters the truth conditions for statements about possibilia. Admitting that no modal model is realistic, one might characterize a transfinite hierarchy of "near-realistic" models of increasingly large size and claim that a sentence about worlds is true if at some point in this hierarchy the sentence is true and remains true from that point onward. But this translation procedure would misrepresent the modal facts: the sentence 'There is an upper limit to the infinite number of individuals that exist in any one possible world' would turn out true, despite the fact that it is possible that there exist arbitrarily many individuals.

IVb. "We're not talking about that!" A less technical objection attacks my claim that talk about possibilities concerns infinitary sentences and modal models that no one until now has bothered to characterize. This objection is not particular to the pluriverse view. One might similarly object to the claim that, all along, we have been talking about linguistic-ersatz worlds, complex set-theoretic constructions out of bits of language; or spatiotemporally isolated concrete worlds; or maximal consistent states of affairs; or a fiction about possible worlds; or any other proposed conception of possibilia.

The answer to all these worries is the same. Various modal concepts in the neighborhood of possible worlds form an interrelated cluster: possible worlds, possible individuals, possible states of affairs, necessity, possibility, and so on. These modal concepts are concepts of properties and relations that play certain roles vis-à-vis each other and vis-à-vis other notions. But this exhausts the nature of the modal concepts; they do not (much) constrain the intrinsic properties of candidate possible worlds and individuals. This is a sort of structuralism, if you like. Modal concepts lay down a structural requirement: our talk of possible worlds and the rest is about any structure that is suited to play the relevant role. I say that pluriverse sentences and realistic models are best suited to play this role, and that is all it takes for talk of worlds to be talk about them.[43]

V. THE PLURIVERSE VIEW COMPARED WITH MODAL FICTIONALISM[44]

My pluriverse view, particularly in its linguistic form, is similar in some ways to Rosen's[45] modal fictionalism. There are, however, important differences between the views that favor the ersatz-pluriverse view.

According to Rosen's fictionalism, talk about possible worlds is like ordinary talk about fictional characters. Just as 'There was a detective named 'Holmes' ' is elliptical for 'According to the Conan-Doyle stories, there was a detective named 'Holmes' ', 'There is a possible world in which donkeys talk' is elliptical for 'According to PW, the fiction of possible worlds, there is a possible world in which donkeys talk'. PW consists of seven postulates, which informally capture Lewis's theory of worlds, together with "an *encyclopaedia*: a list of the nonmodal truths about the intrinsic character of this universe" (*ibid.*, p. 335). The notion of truth according to, or in, a fiction is an undefined primitive of the theory. Rosen goes on to propose an analysis of modality: *P* is necessary if, according to PW, *P*'s worlds translation P^* is true in all worlds; *P* is possible if and only if, according to PW, P^* is true in some world.

---

[43] I favor an analogous response to Kripke's Humphrey objection to counterpart theory; see *Naming and Necessity* (Cambridge: Harvard, 1972), p. 45; and my *Four-Dimensionalism*, pp. 194-96. For other responses to Kripke, see Lewis, *On the Plurality of Worlds*, p. 196; and Hazen, "Counterpart-Theoretic Semantics for Modal Logic," this JOURNAL, LXXVI, 6 (June 1979): 319-38.

[44] Two other views, to which the pluriverse view bears some similarity, are worth mentioning briefly. According to Tony Roy—"In Defense of Linguistic Ersatzism," *Philosophical Studies*, LXXX (1995): 217-42—worlds in which nonactual individuals swap qualitative roles may be distinguished by representing those individuals with "arbitrary names": for example, 'Batman' and 'Robin' in 'Batman plays role R_1 and Robin plays role R_2' and 'Robin plays role R_1 and Batman plays role R_2'. But either (i) these names mean nothing (since they refer to nothing) and so the sentences represent nothing, or (ii) they are implicitly existentially quantified variables (as is suggested by Roy's remarks on p. 226). In case (ii), either (iia) the variables are bound to quantifiers at the beginning of the sentences, in which case the sentences mean the same thing, namely, $\exists x \exists y$ (x plays role R_1 and y plays role R_2), and the worlds have not been distinguished, or (iib) the variables are bound to quantifiers outside the two sentences, in which case the view is starting to look like the pluriverse view. The other view is Fine's reduction of possible-worlds talk, from his postscript to Fine and Arthur Prior, *Worlds, Times and Selves* (London: Duckworth, 1977); see also his "First-Order Modal Theories, Part II—Propositions," *Studia Logica*, XXXIX (1980): 159-202; "First-Order Modal Theories, Part I—Sets," *Noûs*, XV (1981): 117-206; and "First-Order Modal Theories, Part III—Facts," *Synthese*, LIII (1982): 43-122. Roughly, Fine interprets an existential sentence about possibilia like ⌜There is some possible object that ψx⌝ as meaning ⌜possibly, there is an object that ψx⌝, where the existential quantifier in the latter sentence is an actualist quantifier. Fine's view is attractive, but handles quantification over sets of possibilia (see *Worlds, Times and Selves*, pp. 145-48) and sentences attributing cross-world relations less smoothly than the pluriverse view.

[45] "Modal Fictionalism," *Mind*, XCIX (1990): 327-54.

Both fictionalism and the pluriverse theory reinterpret assertions about worlds as assertions about what is true according to a *single* entity describing the entire pluriverse. Rosen could therefore offer something like the pluriverse theory's solution to the problem of descriptive power. This, it seems to me, is fictionalism's most distinctive benefit.[46]

Rosen's fiction differs superficially from pluriverse sentences by being "generational": it describes the pluriverse by (i) describing *one* world (in the encyclopedia), and (ii) supplying a principle that generates new worlds from old. The principle of generation is one of the seven postulates, specifically, the principle of recombination:

> (6e) The totality of universes is closed under a principle of recombination. Roughly: for any collection of objects from any number of universes, there is a single universe containing any number of duplicates of each, provided there is a spacetime large enough to hold them (*op. cit.*, p. 333).

A more important difference concerns the content of the fiction, not just its form. The pluriverse theory entails$_p$ everything about the pluriverse stateable from our vantage point in the actual world, whereas Rosen's fiction entails much less. Instead, Rosen relies on the fact that much more is generally true in a fiction than what is explicitly stated (*op. cit.*, p. 347). Although never explicitly mentioned, it is presumably true in the Sherlock Holmes fiction that Holmes has ten toes.

---

[46] Other than the solution to the problem of descriptive power, fictionalism has no real advantage over linguistic ersatzism. Each appeals to truth according to linguistic entities. The linguistic ersatzer needs an infinitary Lagadonian worldmaking language, but Rosen does as well to construct his encyclopedia. Ersatzism requires modality, whereas Rosen requires his notion of truth in fiction. But below I argue that the notion of truth in fiction must be at least as powerful as modality if fictionalism is to be materially adequate.

My reason for preferring the pluriverse view to Rosen's fictionalism is dissatisfaction with his notion of truth in fiction. If Rosen's fictionalism is to be materially adequate, that notion will need to be very different from the ordinary notion of truth in fiction, and hence will be an obscure, unexplained primitive.

Rosen needs truth-in-fiction because his fiction does not describe the pluriverse exhaustively. But it would be bogus to take this strategy to an extreme. Imagine an extremely thin fiction, consisting of a single sentence 'There are other possible worlds', and a fictionalist who claimed to reduce worlds talk to truth in this thin fiction. The view seems materially inadequate, for the fiction says too little; it seems not, for example, to imply the existence of a world containing blue swans. The thin fictionalist might reply that it is *implicitly* true in the fiction that there is a world with blue swans. But then the notion of truth in fiction becomes unexplained and obscure, for it is not true in any *ordinary* sense that statements about blue swans are true in the thin fiction. Even if we grant the fictionalist *some* notion of implicit truth in fiction, we should not grant him just any such notion.

Although more extreme, my thin fictionalist is a bit like Rosen himself. Rosen claims that:

(8f) According to PW, there is a universe containing blue swans (*op. cit.*, pp. 335-36).

But the truth of (8f) is far from clear. Rosen's idea, presumably, is that (8f) is true in virtue of the encyclopedia and the principle of recombination (principle (6e) above). The encyclopedia will indeed report the existence of swans and blue things in the actual world. But the principle of recombination will not then *logically* imply the existence of a universe with blue swans. There is no way to patch together duplicates of swans and blue things (blue skies, blue shirts, and the like) to arrive at a blue swan. The only hope would be to patch together duplicates of very small actual things—cells, molecules, or maybe even atoms—to form what would be a blue swan. Rosen's fiction PW thus logically entails[47] a certain sentence, BS, about microscopic objects, but it does not logically entail that there is a blue swan.

Consider the conditional statement 'If BS is true, then there exists a blue swan'. Call such conditionals *microreduction laws*; they express necessary truths[48] connecting the macro and micro realms. The problem under consideration would be solved if microreduction laws are true in PW. But since neither Rosen nor anyone else knows exactly what the microreduction laws are, they cannot be built into the fiction explicitly.[49] So Rosen must either admit that it is *not* true in PW that there is a universe with blue swans, in which case modal fictionalism is materially inadequate, or he must say that microreduction laws are implicitly true in PW.

---

[47] Even this is not quite true, for various reasons. One is that (6e) is an incomplete statement of the principle of recombination, for it does not say that the duplicates may be patched together in arbitrary spatial arrangements.

Other necessary truths will need to be implicitly true in PW as well. PW says nothing about the multitude of conceptual, mathematical, logical, and philosophical truths that are usually taken to be necessary. In certain cases where PW is silent (for example, over whether something could be both positively and negatively charged), Rosen argues with some plausibility that there is no modal fact of the matter (*op. cit.*, section 7), but it would be unacceptable to claim this for *all* of these necessary propositions about which PW is silent. The pluriverse theory has no analogous problem: as noted in section IIIc, necessary truths automatically turn out true in every world; this is of course because the modal notion of necessity was assumed in constructing realistic models.

PW is silent, then, about many necessary truths. Therefore, either Rosen's theory is materially inadequate, or, if he insists that these truths *do* hold in all the worlds of the fiction, then 'truth in fiction' has moved far from any ordinary notion of truth in fiction. Given the usual understanding of truth in fiction, surely the Goldbach conjecture (if it is indeed true) is not true in the Holmes stories; but Rosen's fiction PW says no more about mathematics than does the Holmes stories. Perhaps *some* elementary necessary truths could be said to be true in all the worlds of PW in virtue of being common knowledge to the writers and readers of typical fictions. We say, for example, that Sherlock Holmes has a liver, not because this is explicitly mentioned in the Conan Doyle stories but because it is common knowledge that detectives have livers. But many necessary truths are not common knowledge.

In evaluating this objection it is important to distinguish the following two claims:

---

[48] On some views, BS should be augmented with relational facts or facts about the laws of nature.

[49] Compare Lewis, *On the Plurality of Worlds*, pp. 155-56.

> It is true in the fiction PW that every necessary truth is true in all worlds.
>
> For every necessary truth, *N*, it is true in the fiction PW that *N* is true in all worlds.

The first may well be unobjectionable, but it is the second that is required by Rosen's theory, and it is the second that is implausible, if the operative notion of truth in fiction is the ordinary one.

This difficulty for fictionalism emerges in another way, in strengthened form. The encyclopedia and principle of recombination generate many worlds; but Rosen rightly refrains from claiming that all worlds may be thus generated, because of the possibility of fundamental properties that do not actually exist. Indeed, the fiction contains (in postulate (6g)—*ibid.*, p. 333) a claim to the effect that there are worlds beyond those which can be generated in this way. But this means that Rosen's fiction contains no analog of COMPLETENESS (section IIIe), which says that the worlds a pluriverse sentence mentions are *all* of the worlds that exist. Thus, PW does not logically entail any (nontrivial) claims of the form 'There is no world of type *T*' or 'Every world is of type *T*'. The point holds for more complex assertions with universal quantifiers over possibilia, for example: 'Every possible object of type T_1 bears relation *R* to some possible property of type T_2' or 'Every function defined on possible individuals of type T_3 is also of type T_4'. These claims must be accounted for solely by the primitive of truth in fiction, since the fiction is silent on these matters. This goes far beyond the previous point that all necessary truths must be true in every world in the fiction. As mentioned above, there are truths about the pluriverse which are not stateable in the language of necessity and possibility; few of these with universal quantifiers will be logically entailed by PW.

One final problem for fictionalism arises from the fact, just mentioned, that Rosen builds the possibility of properties beyond those which exist in the actual world into his fiction PW, via postulate (6g). Thus it follows from Rosen's analysis of modal concepts that the actual world does not contain all possible properties. This fact about the actual world is likely true, but it should not be built into an analysis of modality, since there may exist possible worlds with all possible properties, and it seems an open epistemic possibility that ours is such a world. Nor should Rosen build into PW the claim that *no* nonactual properties are possible, for then the fiction would be overly bold in the opposite direction: we would have a conceptual guarantee that this *is* the "richest" possible world. Human fictionalists do not know the modal facts, and so do not know what to build into the fiction.

The pluriverse view has no corresponding problem, since pluriverse sentences are constructed as a function of the modal facts. If 'Necessarily, every property actually exists' is an L-true sentence of the modal language then every pluriverse sentence entails_p that @ is the richest possible world; if not, not. Rosen cannot analogously construct PW conditional on the modal facts unless he gives up on the reductive analysis of modality.

Some fictionalists might be happy to give up on reductive analysis, and utilize modal notions in constructing the fiction.[50] If implicit truth-in-fiction were still needed, then fictionalism would require more commitments in ideology than the pluriverse view, which requires only modality. If, on the other hand, the fiction could then be spelled out in perfect detail, fictionalism would become the pluriverse view. Unless the difference between the two views collapses, then, there are reasons to prefer the ersatz pluriverse theory over fictionalism.

THEODORE SIDER

Rutgers University

---

[50] Rosen admits that in some sense 'according to' is a modal primitive (*ibid.*, section 8); the question is whether to admit possibility and necessity in addition.

FAREWELL TO THE DIRECT ARGUMENT[*]

Consider the traditional argument for the incompatibility of moral responsibility and determinism, henceforth the *traditional argument*.

(1) If determinism is true, then no agent could have avoided acting as he did (call this assumption IVD).
(2) An agent is morally responsible for an act he performed only if he could have avoided acting as he did (the *principle of alternative possibilities*, PAP).
(3) Therefore, if determinism is true, then no agent is morally responsible for his acts.

Recently, another argument for the same conclusion has become popular among incompatibilists, namely, one that does not employ the notion of avoidability, but argues "directly" for that conclusion from some general and allegedly uncontroversial assumptions about moral responsibility and determinism. Call that argument the *direct argument*.[1]

An incompatibilist like Peter van Inwagen,[2] who introduced this type of argument into the literature on free will, has done so to show that one could establish the incompatibility of moral responsibility and determinism *without* the "avoidability" or "could-have-done-otherwise" notion of free will (*ibid*., pp. 183-88). Others have found such an argument attractive because of certain worries compatibilists have voiced about the incompatibilist defense of assumption (1) of the traditional argument (IVD). The direct argument, it is claimed, avoids those worries, and hence is more persuasive.[3] Finally, there are incompatibilists who concede that

---

[*] I would like to thank Eleonore Stump, Michael McKenna, Bill Rowe, Eddy Zemach, Hagit Ben Baji, Lior Lavi, Charlotte Katzoff, Yehuda Schnall, and especially Carl Ginet and David Gottlieb for some very helpful comments and discussions on the issues discussed in this article.

[1] I am borrowing the term 'direct argument' from John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (New York: Cambridge, 1998), chapter 6.

[2] *An Essay on Free Will* (New York: Oxford, 1983)

[3] See Michael McKenna, "Source Incompatibilism, Ultimacy, and the Transfer of Non-Responsibility," *American Philosophical Quarterly*, XXXVIII (2001): 37-50, here pp. 40, 47. For those compatibilist worries, see, for example, David Lewis, "Are We Free to Break the Laws?" *Theoria*, XLVII (1981): 113-21; Thomas P. Flint, "Compatibilism and the Argument from Unavoidability," this JOURNAL, LXXXIV, 8 (August 1987): 423-40; and Tomis Kapitan, "Incompatibilism and the Practical Modalities," *Analysis*, LVI (1996): 102-10.

0022-362X/02/9906/316–24

the traditional argument is flawed, because they agree with Harry Frankfurt's[4] criticism of assumption (2) of that argument (PAP). These incompatibilists, however, still hold that determinism is incompatible with moral responsibility. Naturally, they, too, have found the direct argument attractive, precisely because it dispenses with the notion of avoidability—a notion they think is irrelevant to moral responsibility.[5] Although differing in their motivation, all of these parties share the view that the direct argument is an important dialectical tool which can be used to convince compatibilists (or those who are undecided) of the incompatibility of moral responsibility and determinism.

I wish to argue here that the direct argument is not an improvement over the traditional argument. Not only are some standard versions of it invalid; but even if one were to come up with a flawless version, its plausibility would still depend on the traditional assumption that determinism rules out avoidability (IVD). Ultimately, the direct argument cannot do the work its proponents wanted it to do.

I. REFUTATION OF THE DIRECT ARGUMENT

Let us begin with van Inwagen's version of the direct argument (*op. cit.*, pp. 184-88). It is based on the following two rules of inference which he takes to be valid:

(A) $\Box p \vdash NR(p)$
(B) $NR(p), NR(p \supset q) \vdash NR(q)$

where '$\Box$' represents broadly logical necessity, '$\vdash$' stands for 'therefore', and 'NR(p)' abbreviates 'p and no one is (now), or ever has been, morally responsible for the fact that p'.[6] Let us call the second of the two rules the *transfer of nonresponsibility rule*, or *transfer NR* for short. van Inwagen's version of the direct argument may then be formulated as follows.

---

[4] "Alternate Possibilities and Moral Responsibility," this JOURNAL, LXVI, 23 (December 4, 1969): 829-39.

[5] See, for example, Derk Pereboom, *Living without Free Will* (New York: Cambridge, 2001), p. 33; Linda Zagzebski, "Does Libertarian Freedom Require Alternate Possibilities," *Philosophical Perspectives*, XIV (2000): 231-48; Hunt David, "Moral Responsibility and Avoidable Action," *Philosophical Studies*, XCVII (2000): 195-227; and Eleonore Stump, "Alternative Possibilities and Moral Responsibility: The Flicker of Freedom," *The Journal of Ethics*, III (1999): 299-324. These incompatibilists are sometimes referred to as "Causal History Incompatibilists" or as "Source Incompatibilists."

[6] My formulation of van Inwagen's two rules differs slightly from his own. In lieu of 'NR', van Inwagen uses 'N', where his 'N' stands for 'no one is, or ever has been, *even partly responsible* for the fact that...'. Both these differences can be safely ignored, however.

Like van Inwagen, I read the 'is' in the expression 'no one is, or ever has been responsible for the fact that...' as a tensed 'is'. Thus, I view that expression as being equivalent to 'no one is *now*, or ever has been responsible for the fact that...'.

Let 'PAST' stand for a proposition describing the state of the world at some time before the existence of the human race, 'LAWS' stand for the conjunction of the laws of nature, and 'PRESENT' stand for any true proposition about the state of the world now.

(1) Assume now that the thesis of determinism is true.
(2) $\Box$(PAST & LAWS. $\supset$ PRESENT) from 1
(3) $\Box$(PAST $\supset$ (LAWS $\supset$ PRESENT)) from 2, by logic
(4) NR (PAST $\supset$ (LAWS $\supset$ PRESENT)) from 3, by (A)
(5) NR (PAST) assumption
(6) NR (LAWS $\supset$ PRESENT) from 4, 5, by transfer NR
(7) NR (LAWS) assumption
(8) NR (PRESENT) from 6, 7, by transfer NR
(9) Therefore, if determinism is true, then no one is, or ever has been, morally responsible for PRESENT, that is, for any given fact that obtains now.

The direct argument has recently been criticized by Mark Ravizza,[7] who asks us to consider the following scenario, which he calls "Erosion":

> Imagine that Betty plants her explosives in the crevices of a glacier and detonates the charge at T1, causing an avalanche that crushes the enemy fortress at T3. Unbeknownst to Betty and her commanding officers, however, the glacier is gradually melting, shifting, and eroding. Had Betty not placed the dynamite in the crevices, some ice and rocks would have broken free at T2, starting a natural avalanche that would have crushed the enemy's camp at T3 (*ibid.*, pp. 72-73).

Erosion, Ravizza claims, is a counterexample to transfer NR, as the following argument has true premises and a false conclusion.

(1) NR (the glacier was eroding)
(2) NR (the glacier was eroding $\supset$ there is an avalanche that crushes the enemy's base at T_3)
(3) Therefore, NR (there is an avalanche that crushes the enemy's base at T_3)

The conclusion of this argument is false, since Betty acted freely, and is, therefore, responsible for there being an avalanche that crushed the enemy's base. It is not difficult, however, for the proponent of the direct argument to meet this objection. He might point out that in Ravizza's example, Betty is *derivatively* or *indirectly* responsible for the fact that there was an avalanche that crushed the enemy's base. She is derivatively responsible for this fact by virtue of being *directly* responsible for detonating the charge at T_1. The notion of derivative responsibility is parasitic upon the notion of *direct* responsibility. And it is this latter notion of responsibility which ought to be employed in transfer NR. Hence, when properly reformulated, transfer NR withstands Ravizza's objection.

---

[7] "Semi-Compatibilism and the Transfer of Non-Responsibility," *Philosophical Studies*, LXXV (1994): 61-93, here pp. 72-73; see also Fischer and Ravizza, p. 157.

Unfortunately, transfer NR is invalid, even if formulated in terms of direct responsibility. To see this, consider the following example, called "Fate":

> Suppose that Jones murders Smith at T_0 for some selfish reason, and that later on he murders another person, Green, at T_3. Suppose also that the second murder is made possible by the first murder. That is, in the circumstances, the second murder requires the first murder as a causally necessary condition. Finally, suppose that Jones could have avoided murdering both Smith and Green, and believed that he could have avoided to so act.

Assuming that it is now T_1 ($T_1 < T_3$), consider the following argument:

(1) NR (Jones murders Green at T_3)
(2) NR (if Jones murders Green at $T_3 \supset$ Jones murders Smith at T_0)
(3) Therefore, NR (Jones murders Smith at T_0)

I claim that this argument—an instance of transfer NR—is invalid. Premise (1) is true, since no one is now ($=T_1$), or ever has been, directly morally responsible for a murder that has not been committed yet.[8] And since we assumed that Jones's murdering Smith at

---

[8] One may argue that my claim that no one is now ($=T_1$), or ever has been, directly morally responsible for a murder that has not been committed yet, presupposes the following more general assumption:

(a) No one is now, or ever has been, morally responsible for an event that has not occurred yet.

which is objectionable. For if someone *S* initiates at T_1 (or prior to T_1) a causal chain inevitably leading to an event *E* occurring at time T_3 in the future, then *S* may be deemed morally responsible for that event even before it occurs.

Reply: I find this objection unconvincing for two reasons. Contrary to the

T_0 is a causally necessary condition of his murdering Green at T_3, (2) is true as well. But notice that the conclusion is false, since we assumed that Jones could have avoided killing Smith.[9]

In a recent article, Ted Warfield[10] has tried to defend the direct argument by replacing transfer NR with another transfer rule, which shall be called "transfer NRW":

$\mathrm{NR}(p)$, p entails $q \vdash \mathrm{NR}(q)$

But it is not difficult, however, to refute this rule as well. Letting

p = Peter commits suicide at T_2
q = Peter did not commit suicide before T_2
now = T_1, and T_1 occurs shortly before T_2

---

critic, I hold that moral responsibility is a backward-looking notion. Second, my counterexample to transfer NR does not involve the kind of scenario described by the critic. In "Fate," neither Green's death nor Jones's act of murdering Green is an outcome of a causal chain initiated by Jones or someone else at T_1 (or at a time earlier than T_1.) Thus, "Fate" is compatible with the following restricted version of (a):

> (a′) No one is now, or ever has been, morally responsible for an event E that has not occurred yet, except in cases where someone is responsible for there existing in the present a causal chain inevitably leading to E.

[9] This example brings out nicely some of the asymmetries between the notion of avoidability and that of moral responsibility. For example, although an agent's performance of some action in the past is *now* unavoidable, he may nevertheless *now* be morally responsible for it. And although an agent cannot be *now* responsible for an act he has not performed yet, that act might *now* be avoidable. Pace van Inwagen, pp. 93-104, these differences show that the transfer NR rule and Inwagen's rule Beta cannot be treated on a par.

According to Fischer and Ravizza, transfer NR is also refuted by Frankfurt-style counterexamples to the principle of alternative possibilities (PAP) (*op. cit.*, pp. 157-58). This claim is further developed by Fischer in his "The Transfer of Non-Responsibility," forthcoming in Joseph Keim Campbell, Michael O'Rourke, and David Shier, eds., *Freedom and Determinism: Topics in Contemporary Philosophy Series, Volume II* (New York: Seven Bridges, 2002). Since Frankfurt-style counterexamples to PAP are disputable, however, I find this line of argument against transfer NR unconvincing. For a criticism of the Frankfurt examples, see my "Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities," *Philosophical Review*, CIV (1995): 247-61; Carl Ginet, "In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Argument Convincing," *Philosophical Perspectives*, X (1996): 403-17; and Robert Kane, *The Significance of Free Will* (New York: Oxford, 1996), pp. 142-44.

[10] See his "Determinism and Moral Responsibility Are Incompatible," *Philosophical Topics*, XXIV (1996): 215-26.

we easily obtain a counterexample to it. The reasons for the invalidity of transfer NRW are similar to those mentioned in connection with Fate.[11] We may call this example "Suicide."[12]

In concluding this section, let us take a brief look at one additional argument for the incompatibility of moral responsibility and determinism sometimes employed by incompatibilists—the *ultimacy argument*:

(1) If determinism is true, then no one is the ultimate source of one's acts.
(2) One is morally responsible for one's acts only if one is their ultimate source.
(3) Therefore, if determinism is true, no one is morally responsible for one's acts.[13]

Although not exactly a "direct-type" argument, it may be instructive to examine this argument for two reasons: (i) like the direct argument, it does not employ the notion of avoidability; (ii) its first assumption (if it is going to be of interest to nonincompatibilists) depends for its plausibility on the soundness of the following argument for it, an argument analogous to the direct argument.

(1a) Assume that the thesis of determinism is true
(2a) $\Box$(PAST & LAWS. $\supset$ PRESENT)

---

[11] In response to this objection to transfer NRW, one might claim that, if one were to use van Inwagen's original formulation of the NR-operator, to wit, 'no one is, or ever has been, even *partly* morally responsible for the fact that', then transfer NRW would not be open to my counterexample. For van Inwagen defines the notion of partial responsibility in terms of someone (or some group of human beings) being responsible for the logical consequences of the fact that p (*op. cit.*, p 243, note 28). Consequently, '$N(p)$' (p = Peter commits suicide at T_2) in my counterexample would turn out to be false; for 'Peter commits suicide at T_2' entails 'Peter was born before T_2'; and at T_1, someone is responsible for the fact that Peter was born before T_2, namely, his parents.

Reply: first, van Inwagen's account of partial responsibility is too strong even on his own terms. According to him, the notion of partial responsibility is weaker than the notion of responsibility in the sense that being responsible for some fact implies being partially responsible for it, and not vice versa. But then no one could be responsible for the fact that he committed suicide. For the latter fact entails that he was born, and no one can be held responsible for the fact that he was born. Second, van Inwagen's use of 'even' in his formulation of NR(p) in terms of 'p and no one is, or ever has been, *even* partly responsible for the fact that p' strongly suggests that he wishes the direct argument to hold not only for the notion of partial responsibility but also for the notion of responsibility itself. And the counterexample in the text does refute transfer NRW when formulated just in terms of responsibility.

[12] For a different, but weaker counterexample to Warfield's rule, see Eleonore Stump and Fischer, "Transfer Principles and Moral Responsibility," *Philosophical Perspectives*, XIV (2000): 47-55, here pp. 49-50. I say "weaker," since like Ravizza's counterexample to van Inwagen's rule, it involves a case of derivative responsibility.

It may be of interest to note that "Suicide" is also a counterexample to van Inwagen's transfer NR rule.

[13] For this formulation of the argument, see McKenna, p. 40; but he does not endorse this argument.

(3a) NU (PAST & LAWS. $\supset$ PRESENT)
(4a) NU (PAST & LAWS)
(5a) NU (PRESENT)
(6a) Therefore, determinism implies that no one is, or, ever has been, the ultimate source of any given fact that obtains now.

This argument, in turn, assumes the validity of a transfer rule analogous to transfer NR called *transfer NU.*

$$\mathrm{NU}(p),\ \mathrm{NU}(p \supset q) \vdash \mathrm{NU}(q)$$

where '$\mathrm{NU}(p)$' stands for 'p and no one is, or ever has been the ultimate source of its being the case that p'. As one can easily see, however, both "Fate" and "Suicide" constitute counterexamples to transfer NU. Hence, the ultimacy argument seems to be subject to the same kind of problems as is the direct argument.

II. A STRONGER OBJECTION

At this point, proponents of the direct argument may argue that all that examples such as "Fate" and "Suicide" show is that transfer NR needs to be suitably restricted. And, indeed, various ways of restricting it have been proposed.[14] I do not think, however, that such a move is likely to save the direct argument.

To see this, consider again Jones, who deliberately murders Smith for some selfish reason, knowing very well that in doing so he is acting immorally. Suppose further that Jones lives in a deterministic world. Is Jones morally blameworthy for what he did? Proponents of the direct argument will want to answer this question negatively. But suppose that their opponent disagrees, claiming that Jones *is* blameworthy, since he knew that he was acting wrongly and *could* have avoided acting as he did. What can the defender of the direct argument say in reply? It seems that his only and obvious response would be to say that, in the deterministic world under consideration, Jones could *not* have avoided killing Smith. Otherwise, there would be no reason to regard the improved version of transfer NR as valid, as it would be rendered invalid by the deterministic scenario under consideration.[15] But notice that the dialectical situation has changed. For now it is clear that the validity of the improved version of transfer NR (*transfer NR**) requires that determinism rules out avoidability. In response to this objection, the proponent of the direct argument may argue that he is more sure of the validity of transfer NR* than he is of the assumption that

---

[14] For example, Ginet has suggested the following improvement of transfer NR:

$\mathrm{NR}(p),\ \mathrm{NR}(p \supset q) \vdash \mathrm{NR}(q)$, where q describes an event or state of affairs that occurs later than that described by p.

See Ginet, "Libertarianism," forthcoming in Dean Zimmerman and Michael Loux, eds., *Oxford Handbook of Metaphysics* (New York. Oxford, 2002). For other attempts to revise transfer NR, see McKenna, p. 45; and Stump, "Control and Causal Determinism," forthcoming in Sarah Buss and Lee Overtone, eds., *The Contours of Agency. Essays in Honor of Harry Frankfurt* (Cambridge: MIT, 2002).

> (M) Someone who knowingly acted in a morally wrong way, and believed correctly that he could have avoided acting as he did, is blameworthy for his act.

and, therefore, he does not see a reason *not* to use transfer NR*.[16] But such a reply would be implausible. Surely, (M), which is explicative of the very notion of blameworthiness, enjoys a degree of certainty that does not fall below that of transfer NR*. Hence, (M) can certainly be viewed as a defeater of that principle.

III. THREE MORALS AND A FURTHER OBJECTION

What are the implications of the above result? First, there is an important logical implication: the direct argument (or a suitably revised version of it) depends for its plausibility on the assumption (IVD) of the incompatibility of determinism with the freedom to avoid acting as one did. Hence, it cannot do the work its proponents wanted it to do. Second, the direct argument should not be of concern to compatibilists, since they reject the assumption that in a deterministic world the agent cannot act otherwise. Third, proponents of the direct argument who reject PAP cannot ignore the question of the avoidability of the agent's act, contrary to their initial intention. What is more, they are now also called upon to answer the following question: Given that, on their view, unavoidability does not exonerate the agent, why is the assumption of the unavoidability of the agent's act in a deterministic universe so crucial to the success of the direct argument?[17]

In a way, all this is not so surprising. For the hope of proponents of the direct argument to be able to prove that determinism excludes moral responsibility without the notion of avoidability was unrealistic in the first place. To see this, it is enough to consider one central assumption of the direct argument, to wit, that no one is morally responsible for the laws of nature. Why should one accept it if not for the fact that no one could have prevented them from obtaining? It will not do for the proponent of the direct argument to reply that this is not the reason why we are not morally responsible for the laws of nature; and that the true reason is that no human being brought them about. This reply assumes that one's not being causally responsible for a fact is sufficient for one's not being morally responsible for that fact. But this assumption is false. We often are morally responsible for facts or events for which we bear *no* causal responsibility. I might be morally responsible for the event of Smith's being bitten by a cobra even if I did not bring about this event, given that I knew that it was about to happen and could have prevented it. So the idea underlying the direct argument—to establish the incompatibility of moral responsibility with determinism without employing the notion of avoidability—is misguided.

---

[15] Note that this objection does not assume that avoidability is a necessary condition for moral responsibility (PAP). What it does assume is only that avoidability in conjunction with some other assumptions is *sufficient* for responsibility.

[16] I would like to thank Ginet for stressing the need to address this objection.

[17] This question is especially pertinent if one does not define determinism as the view that every event is caused, but as the view that every event has a *nomically sufficient* condition. (Roughly speaking, an event E occurring at time T has a *nomically sufficient* condition if and only if the proposition that E occurs at T is entailed by the laws of nature and some proposition p that expresses the state of the world at some time prior to T)

In the light of all these criticisms, I propose that we cease to see in the direct argument an argument that can usefully advance the debate between incompatibilists and their opponents.

DAVID WIDERKER

Bar-Ilan University

COMMENTS AND CRITICISM

STANLEY AND WILLIAMSON ON KNOWING HOW

In "Knowing How," Jason Stanley and Timothy Williamson[1] propose an account of knowing how which makes it a species of propositional knowledge, or knowing that. Consider an attribution of knowledge of how to perform an action like this:

(1) Hannah knows how to ride a bicycle.

On their account, this attribution is true just in case there is some way *w* which is a way for Hannah to ride a bicycle, Hannah knows that *w* is a way for her to ride a bicycle, and "Hannah entertains [the proposition that *w* is a way for her to ride a bicycle] under a practical mode of presentation" (430).[2] *Ways* are properties of token events, and the thrust of their proposal is that to know how to *F* is to possess the propositional knowledge of what to do in order to *F*, or to know of some way to *F*. The motivation of the requirement that Hannah entertain the relevant proposition "under a practical mode of presentation" is illustrated by the following example: "Suppose that Hannah does not know how to ride a bicycle. Susan points to John, who is riding a bicycle, and says [to Hannah], 'That is a way for you to ride a bicycle'. Suppose that the way in which John is riding his bicycle is in fact a way for Hannah to ride a bicycle" (428). Letting the demonstrative 'that way' denote John's way of riding a bicycle, it seems true that

(2) Hannah knows that that way is a way for her to ride a bicycle.

Nevertheless (1) is false. The difference between (2) and (1), according to Stanley and Williamson, is that while both ascribe to Hannah knowledge of the same proposition,[3] in (2) "knowledge of the proposition is ascribed to Hannah under a demonstrative presentation, [while in (1)], knowledge of that proposition is ascribed to Hannah

---

[1] This JOURNAL, XCVIII, 8 (August 2001): 411-44.

[2] As they note, this is to incorporate the proposition's mode of presentation into the truth conditions of (1). If one instead treats the mode of presentation as of only pragmatic relevance, then a use of (1) pragmatically conveys that she entertains the proposition under a practical mode of presentation.

[3] They employ the Russellian notion of a proposition, on which the relevant proposition consists of an ordered sequence of a way for someone to ride a bicycle and Hannah.

0022-362X/02/9906/325–28

under a different mode of presentation, what we call a *practical* mode of presentation" (429).

Stanley and Williamson admit to being unable to give a nontrivial characterization of what it is for a way to perform an action to be presented under a practical mode, though they do say that "thinking of a way under a practical mode of presentation undoubtedly entails the possession of certain complex dispositions" (429), and that "[it] is simply a feature of certain kinds of propositional knowledge that possession of it is related in complex ways to dispositional states" (430). One might be tempted to say that the reason (2) can be true without Hannah's knowing how to ride a bicycle is that she can know that that way is a way for her to ride a bicycle without being able to instantiate that way herself (recall that ways are properties). But Stanley and Williamson correctly note that one can know how to do something without having the ability to do it, as in the case of a pianist who knows how to play the piano, even though an accident in which he has lost his hands has deprived him of the ability to play it. But what this suggests, then, is that the trouble with (2) is that while Hannah knows that that way is a way for her to ride a bicycle, she does not know how to instantiate that way herself. More generally, it suggests that what Stanley and Williamson's account of what it is for someone to know how to F really comes to, with its talk of entertaining a proposition under a practical mode of presentation, is that someone knows how to F just in case he knows, of some way w, that w is a way for him to F, and he knows how to instantiate w himself (and thereby F). For suppose that someone could know, of some way w, that w is a way for him to F, and could entertain that proposition under a practical mode of presentation, and yet not know how to instantiate w himself. Then it would seem that he could, on Stanley and Williamson's account, know how to F without being able to F (since he would not know how to instantiate w and hence be unable to instantiate w and thereby F[4]), even though there is no impediment, handicap, or other barrier to his being able to F (as there is in the case of the maimed pianist). Yet in invoking their account of knowing how to criticize a reply of David Lewis's to Frank Jackson's well-known "knowledge argument," they reject as "absurd" the suggestion that Jackson's imaginary scientist Mary might know how to imagine an experience of red while still in her black and white room, and yet be unable to imagine an experience of red (443). It appears, then, that entertaining the proposition that w is a way for one to F under a practical mode of presentation involves knowing how to instantiate w one's self. But if so, then Stanley and Williamson's account of knowing how appeals to the very notion it seeks to explicate.

---

[4] Against this it might be said that there are examples—like a baby's sucking, to adapt one they take from Paul Ziff (440)—of actions people are able to perform without knowing how to perform, and that instantiating w is such an action. See below for a response

It is true that we sometimes seem willing to attribute knowledge of how to accomplish a complex action or task to someone who does not know how to do what is required to actually accomplish it. For instance, Paul Revere may know that the way to alert the populace to the imminent arrival of the British is to get on a horse and ride to every Middlesex village and farm. But what if he does not know how to ride a horse? Does he know how to alert the populace (he certainly lacks the ability)? We might be willing to say that he does, just as we might be willing to say that a political candidate knows how to win an election, namely, by raising enough money to run commercials portraying his opponent as a philandering draft-dodger, even though he does not himself know how to raise that much money.[5] But as the actions in question become simpler, approaching so-called basic actions, our inclination to attribute knowledge of how to do something in the absence of knowing how to bring it off dissolves. I may know that wiggling my ears is a way for me to annoy Diane, and so know how to annoy her, without knowing how to wiggle my ears. But if I know that what I have just seen Tom do in wiggling his ears is a way for me to wiggle my ears too, and yet do not know how to instantiate that way myself, I do not know how to wiggle my ears.

Stanley and Williamson begin their paper by criticizing an argument of Gilbert Ryle's[6] against construing knowing how as a species of knowing that. Ryle argued that, if performing an action requires one to know how to perform it, and if knowing how to perform an action requires one to entertain or contemplate a proposition, then a regress would ensue, since entertaining or contemplating the proposition would itself be an action one would have to know how to perform, which would require one to entertain or contemplate a distinct proposition, and so on. Stanley and Williamson rightly note that this argument requires the action initially at issue to be an *intentional* action, and that the entertainment or contemplation of a proposition need not be taken to be an intentional action; so the regress is avoided. It is ironic then that their own account of knowing how as a species of knowing that seems vulnerable to an objection similar in spirit to Ryle's. Suppose that to *F* is to perform an intentional action. Then, as Stanley and Williamson acknowledge (442-43), one must know how to *F* in order to *F*. If knowing how to *F* is to know, of some way *w*, that *w* is a way for one to *F* and to know how to instantiate *w* one's self, then either knowing how to instantiate *w* one's self requires knowing, of some way *w′*, that *w′* is a way to instantiate *w* one's self and knowing how to instantiate *w′* one's self; or it does not. If it does, then we move on to the next step of a regress. If it does not, then we have an instance of knowing how which is not, contrary to Stanley and Williamson's account, an instance of knowing that. And it is not open to them to object, as they did to Ryle's argument, that in doing what one has to do in order to *F*—that is, in instantiating *w* one's self—one need not be thought of as performing an intentional action (and hence not an action one needs to know how to perform); for in doing so one *F*s, and thereby performs an action which by assumption is intentional.

---

[5] Stanley and Williamson's example (which they attribute to Jeff King) of the ski instructor who knows how to perform a complex trick without being able to perform it herself would seem to fall into this category.

[6] Presented in "Knowing How and Knowing That," in his *Collected Papers, Volume 2* (New York: Barnes and Noble, 1971), pp. 212-25, and in *The Concept of Mind* (Chicago: University Press, 1949), chapter 2.

JOHN KOETHE

University of Wisconsin/Milwaukee

BOOK REVIEWS

The Discovery of Things: Aristotle's Categories *and Their Context.* WOLFGANG-RAINER MANN. Princeton: Princeton University Press, 2000. xii + 231 p. Cloth $41.50.

The *Categories* occupies a prominent but curious place in our understanding of Aristotle. Although it comes first in the traditional order of his works, it begins without a word of explanation. While the *Metaphysics* famously begins by declaring that all human beings by nature desire to know, the *Categories* begins: "When things have only a name in common, but the definition of being which corresponds to the name is different, they are called homonymous." The rest of the first chapter is given over to definitions of synonymy and paronymy: synonyms share a common name and the corresponding definition; something which takes its name from something else but with a change in ending—as the grammarian takes his name from grammar—is paronymous. These arid distinctions would make a discouraging beginning for any book, let alone one meant to introduce the whole of philosophy. The fact that the traditional order was not Aristotle's own doing helps only a little. Almost from the beginning there has been disagreement over what the book is about and therefore what it should be called. Part of the problem is seeing how its parts fit together to form a unified whole.

The view that prevailed in antiquity concentrates on the central chapters, 4-9, the so-called *Praedicamenta*, that is, categories or predicates. According to it, the subject of the book is ultimate kinds of predicate: quality, quantity, and so on. On this view, the *Categories* belongs at the beginning of the *Organon*, the collection of logical works put together by Aristotle's ancient editors, because it is about terms, the elementary constituents of propositions. On this view, propositions are studied in their turn by the *De interpretatione*, which prepares the way for the treatment of argument in the *Prior Analytics*, varieties of which are then tackled in the *Posterior Analytics*, *Topics*, and *Sophistical Refutations*. The *Organon* (instrument) was so named and put first because the ancient editors regarded logic as an instrument with which to pursue philosophical inquiry rather than a substantive object of inquiry itself.

Although Aristotle appears to have shared this view of logic, the plan that the editors imposed on the *Organon* is much too tidy, especially so far as it concerns the *Categories*, and there were other

0022-362X/02/9906/329–32

views in antiquity reflected in the alternative titles that have come down to us. One, roughly *Preliminaries to the Topics*, retains the connection with argument, but plausibly brings the work into relation with the *Topics*, with which it has many points of contact, rather than the *Prior Analytics*, with which it does not. Another, *On the Genera of Being*, takes its subject to be metaphysical, and the prevailing tendency in recent times has been to mine the *Categories* for insights into Aristotle's early views of substance before turning to the mature reflections of the *Metaphysics*. Of course, the choice between logic and metaphysics may be too narrow. The *Topics* is a handbook of dialectic. Arguments of the kind it treats aim to establish, or more commonly overthrow, definitions. To this end, they presuppose a framework of very general assumptions about being or essence, which it is the task of a definition to specify. The idea that the *Categories* is most closely related to the *Topics*, then, need not make its use as a source for Aristotle's early metaphysical thought illegitimate, for, according to the view at issue, it was something like an introduction to the metaphysical foundations of dialectic.

Wolfgang-Rainer Mann's account of the *Categories* is squarely in this metaphysical tradition. But while many interpreters have been chiefly interested in tracing the path from the *Categories* to the *Metaphysics*, his focus is on the distance Aristotle covered on his way to the position of the *Categories*. One of his principal achievements is an interpretation which, by uncovering the roots of that position in Aristotle's critical engagement with middle-period Platonism, permits us to see how at least the first two parts of the *Categories* (1-9) form a unity. According to this interpretation, homonymy and the other so-called "-onymies" are here to replace Platonic eponymy—which recognizes only a single relation between particular *F*s and the form *F* after which they are named—with several relations expressed in a variety of ways. (Note that according to the relevant way of speaking, if we say, for example, 'Paul is selfish', both 'Paul' and 'selfish' become names of Paul.)

To see this, we need to attend less to horizontal cases, in which items on a level with each other are called by the same name—for example, a cow and a horse 'animal', a pair of white-colored things 'white', the side of a river and a financial institution 'bank'—and more to vertical cases, in which particulars are named after the universals under which they stand. According to this proposal, it is possible to be named *after* something homonymously or synonymously. This cow and the species cow are called 'cow' synonymously; this white thing and the color white and are called 'white' homonymously. If this is right, horizontal cases in which the account of why *x* bears the name *N* has nothing to do with the account of why *y* does—chance homonyms as Aristotle elsewhere calls them—though the clearest instances of homonymy, are misleading. Although what it is to be (the color) white and to be white (in color) are different, the relation between them is crucial.

The "-onymies," then, prepare the way for the recognition of two relations behind eponymy, where middle-period Platonism appears to have seen only a single undifferentiated relation of participation between particulars and universals. And in chapter 2, instead of two sorts of things joined by one relation, participants and the forms in which they share, we now have four kinds of things connected by two relations: particular things or objects, the kinds to which they belong, the attributes which inhere in them, and the kinds to which these attributes belong. The break with Platonism which has long been seen here is underscored when Aristotle goes on in chapter 5 to award the title "primary substance" to the particular objects, leaving their kinds as "secondary substances."

For Mann's purposes, the most important consequence is that there is now something for ordinary objects to be as well as attributes for them to have. The second and longest part of his book is devoted to establishing just how much of a departure this apparently obvious and unexceptionable idea is from what went before by developing an account of middle-period Platonic metaphysics from which anything corresponding to ordinary objects as we usually understand them is conspicuously absent. This is the moral expressed by the title of Mann's book and restated in it: "Plato does not recognize things as things"; "things do not show up as things" before the *Categories*. The interpretation he develops to this end is subtle and wide ranging. It is possible to mention only a few of its main features here.

In agreement with illustrious predecessors like John Burnet and David Furley, Mann emphasizes Plato's affinities with the pre-Socratics, especially Anaxagoras. Instead of objects which are, say, cows, on the one hand and, have, say, the attribute brown, on the other, we should instead think of mixtures made up of shares in more elemental beings, namely, Anaxagorean stuffs or Platonic forms. The nature of the mixtures, the beings in which they share and the relation of sharing in which they stand to them are of course very different. But in each theory, though called by the names *A*, *B*, and *C* after the elemental beings *A*, *B*, and *C* in which they participate, the participants are not fully or really *A*, *B*, or *C* in the same way.

Plato's preferred way of drawing this contrast is to oppose *becoming F*, of which sensible particulars are capable, to *being F*, of which they are not. In one of the book's most interesting sections, Mann is able to throw a good deal of light on this puzzling feature of Plato's thought by tracing it back to a prephilosophical use of 'become' which he calls the funerary use because of its prominence in funeral orations. The idea is that to become *F* in this sense is to display, perhaps temporarily or occasionally, the characteristics of an *F*, which it is possible to do without really being *F*. Mann's interpretation is also unusual in the attention it pays to the position assigned by Plato to the "late learners" in the *Sophist*, who are so called because, like those new to the use of words, they think it proper to say of an *x* only that it is an *x*. Applied to particulars, this yields a position at the opposite pole to the Anaxagorean Platonism of the middle dialogues: it allows things to be precisely what they *are* and to be (or become) nothing else at all. On this view, the *Categories* finds a middle way between the two poles which allows things both to be what they are and, in another way, to be many other things as well by having them as attributes.

In the third and last part of the book, Mann returns to the *Categories* in order to place it in the Platonic context that he has reconstructed for it in the second part and to confirm the path-breaking character of its doctrines. He has a fascinating story to tell, in the course of which he connects and makes sense of issues of central importance in Aristotelian, Platonic, and pre-Socratic philosophy. It is not the only story that can be told. Others have and will make different deductions from the clues. But it is fresh and illuminating and deserves to be widely read. What is more, apart from its considerable merits as an interpretation of the *Categories* and its background, *The Discovery of Things* is an excellent read, well served by its unusual structure and an exceptionally clear and graceful style. For this reason it should be of interest not only to seasoned students of Plato and Aristotle, but also to those cutting their teeth on ancient philosophy.

JAMES ALLEN

University of Pittsburgh

New and Noteworthy from Cambridge

The Play of Character in Plato's Dialogues
Ruby Blondell
0-521-79300-9 Hardback $65 00

The Ethical Dimensions of the Biological and Health Sciences
A Textbook with Case Studies
Second Edition
Ruth Ellen Bulger,
Elizabeth Heitman, and
Stanley Joe Reiser, Editors
0-521-81053-1 Hardback $90.00
0-521-00886-7 Paperback $30 00

The Logic of Concept Expansion
Meir Buzaglo
0-521-80762-X Hardback $55.00

Jacques Derrida and the Humanities
A Critical Reader
Tom Cohen
0-521-62370-7 Hardback $70 00
0-521-62565-3 Paperback $25.00

Nature, God and Humanity
Envisioning an Ethics of Nature
Richard L Fern
0-521-81122-8 Hardback $65.00
0-521-00970-7 Paperback $23.00

Liberal Pluralism
The Implications of Value Pluralism for Political Theory and Practice
William A. Galston
0-521-81304-2 Hardback $55 00
0-521-01249-X Paperback $19 00

Practical Reasoning in a Social World
How We Act Together
Keith Graham
0-521-80378-0 Hardback $55.00

Karl Popper— The Formative Years, 1902-1945
Politics and Philosophy in Interwar Vienna
Malachi Haim Hacohen
0-521-89055-1 Paperback $25.00

Symbols of Jesus
A Christology of Symbolic Engagement
Robert Cummings Neville
0-521-80787-5 Hardback $65.00
0-521-00353-9 Paperback $23 00

Narrative, Religion and Science
Fundamentalism versus Irony, 1700-1999
Stephen Prickett
0-521-81136-8 Hardback $65.00
0-521-00983-9 Paperback $23.00

Living Together and Christian Ethics
Adrian Thatcher
New Studies in Christian Ethics 21
0-521-80204-0 Hardback $65.00
0-521-00955-3 Paperback $23.00

Herodotus in Context
Ethnography, Science and the Art of Persuasion
Rosalind Thomas
0-521-01241-4 Paperback $25.00

Epicurus and Democritean Ethics
An Archaeology of Ataraxia
James Warren
Cambridge Classical Studies
0-521-81369-7 Hardback $65 00

Available in bookstores or from
CAMBRIDGE
UNIVERSITY PRESS
800-872-7423
us.cambridge.org/philosophy

new in philosophy

The Illusion of Conscious Will

Daniel M. Wegner

"This book will serve as the foundation for an untold number of hot debates on who is in charge of our personal destinies." — Michael S. Gazzaniga, Dartmouth College

A Bradford Book • 408 pp. $34.95

Gadamer's Century

Essays in Honor of Hans-Georg Gadamer

edited by Jeff Malpas, Ulrich Arnswald, and Jens Kertscher

A wide-ranging collection of philosophical essays in honor of Hans-Georg Gadamer.

Studies in Contemporary German Social Thought • 336 pp. $25 paper

Furnishing the Mind

Concepts and Their Perceptual Basis

Jesse J. Prinz

"The most important work on concepts to have been written since Locke's Essay." — Fiona Cowie, California Institute of Technology

Representation and Mind series

A Bradford Book • 375 pp., 12 illus. $40

Vision and Mind

Selected Readings in the Philosophy of Perception

edited by Alva Noë and Evan T. Thompson

A collection of works, many of them classics, on the orthodox view of visual perception.

A Bradford Book

605 pp., 39 illus. $38 paper

What Genes *Can't* Do

Lenny Moss

A historical and critical analysis of the concept of the gene that attempts to provide new perspectives and metaphors for the transformation of biology and its philosophy.

A Bradford Book • Basic Bioethics series

258 pp., 2 illus. $34.95

Terms and Truth

Reference Direct and Anaphoric

Alan Berger

"A major work. Berger offers an elegant examination of issues that have been in controversy for the last forty years." — Gilbert Harman, Princeton University

A Bradford Book • 232 pp. $30

Seeing Double

Shared Identities in Physics, Philosophy, and Literature

Peter Pesic

"Drawing on philosophy, literature, and physics in accessible prose, Peter Pesic illuminates the meaning of unique personhood. A challenging and civilizing tour-de-force." — Gerald Holton, Harvard University

192 pp., 11 illus. $24.95

Belief's Own Ethics

Jonathan E. Adler

"An original and important contribution to the ethics-of-belief tradition in epistemology." — John Woods, University of British Columbia

A Bradford Book • 403 pp. $40

Global Justice and Transnational Politics

edited by Pablo De Greiff and Ciaran Cronin

Essays exploring the prospects for transnational democracy in a world of increasing globalization.

328 pp. $22.95 paper

now in paperback

Beyond Hegel and Nietzsche

Philosophy, Culture, and Agency

Elliot L. Jurist

"There is a great deal to learn from this fine book." — Simon Critchley, University of Essex and College International de Philosophie

Studies in Contemporary German Social Thought • 369 pp $22 paper

To order call **800-405-1619**

Prices subject to change without notice

from The MIT Press

The Letters of George Santayana Book Three, 1921-1927

The Works of George Santayana Volume V, Book Three

George Santayana
edited and with an introduction by
William G. Holzberger

The third of eight books of the correspondence of George Santayana.
600 pp $55

The Iconic Logic of Peirce's Graphs

Sun-Joo Shin

"This book reports extremely exciting new results which have major importance for the understanding of graphical logics and important implications outside logic for understanding cognition." — Keith Stenning, University of Edinburgh

A Bradford Book
207 pp., 180 illus. $34.95

now in paperback

Dynamics in Action

Intentional Behavior as a Complex System

Alicia Juarrero

"This work seems to me to be a major new creative contribution to the field of action theory, one that is urgently needed." — W. Norris Clarke, *International Philosophical Quarterly*

A Bradford Book
300 pp., 7 illus. $22 paper

now in paperback

Bioethics

Ancient Themes in Contemporary Issues

edited by Mark G. Kuczewski
and Ronald Polansky

"A must read for those in bioethics." — Patricia Werhane, University of Virginia

A Bradford Book • Basic Bioethics series
320 pp. $24 paper

The Contours of Agency

Essays on Themes from Harry Frankfurt

edited by Sarah Buss
and Lee Overton

"Promises to be one of the most important and most widely read anthologies in moral philosophy and the philosophy of mind in many years." — Gideon Rosen, Princeton University

A Bradford Book • 385 pp., 1 illus $45

Self-Trust and Reproductive Autonomy

Carolyn McLeod

"McLeod's book will help the general reader understand important moral issues surrounding miscarriage, infertility treatment, and prenatal diagnosis." — Rosemarie Tong, University of North Carolina at Charlotte

A Bradford Book • Basic Bioethics series
200 pp. $29.95

Divided Natures

French Contributions to Political Ecology

Kerry H. Whiteside

"This is an outstanding contribution to the field of environmental philosophy/politics. Nothing like it exists in the English-language literature." — Michael Zimmerman, Tulane University

336 pp. $24.95 paper

now in paperback

On the Pragmatics of Social Interaction

Preliminary Studies in the Theory of Communicative Action

Jürgen Habermas
translated by Barbara Fultner

Offers an excellent introduction to Habermas' work, focusing on the germs of the ideas he would elaborate in *The Theory of Communicative Action.*

Studies in Contemporary German Social Thought • 216 pp. $18 paper

http://mitpress.mit.edu

New from Princeton

Imperfect Garden
The Legacy of Humanism
Tzvetan Todorov

Available in English for the first time, *Imperfect Garden* is both an approachable intellectual history and a bracing treatise on how we should understand and experience our lives. In it, one of France's most prominent intellectuals explores the foundations, limits, and possibilities of humanist thinking. Tzvetan Todorov seeks an answer to modernity's fundamental challenge: how to maintain our hard-won liberty without paying too dearly in social ties, common values, and a coherent and responsible sense of self.

"*Imperfect Garden* is a splendid defense of moderate humanism. Tzvetan Todorov carefully reads older French thinkers—Montaigne, Descartes, Montesquieu, and Rousseau—to find a solution to modernity's predicaments."—Thomas Pavel
Cloth $29.95 ISBN 0-691-01047-1

New in paperback
What Makes Us Think?
A Neuroscientist and a Philosopher Argue about Ethics, Human Nature, and the Brain
Jean-Pierre Changeux and Paul Ricoeur
Translated by M. B. DeBevoise

"The two thinkers lock horns about the grandest of issues: the nature of mind, brain, religion, art, morality, and consciousness . The broad terms of the debate and the erudition of the debaters provide considerable insight into how certain neuroscientists and philosophers attack enigmas that first occupied the ancient Greeks."—Howard Gardner, *Chronicle of Higher Education*
Paper $17.95 ISBN 0-691-09285-0

Princeton University Press

PAST AND FUTURE

NEW FROM CHICAGO

THE MOMENT OF COMPLEXITY

Emerging Network Culture

Mark C. Taylor

"A profoundly original work. In remarkable and insightful ways, Mark Taylor traces an entirely new way to view the evolution of our culture, detailing how information theory and the scientific concept of complexity can be used to understand recent developments in the arts and humanities. This book will ultimately be seen as a classic." —John L. Casti, author of *Gödel: A Life of Logic, the Mind, and Mathematics*

CLOTH $32.00

MEMORIES OF ODYSSEUS

Frontier Tales From Ancient Greece

François Hartog

Translated by Janet Lloyd

Drawing on a remarkable range of authors and texts, ancient and modern, François Hartog looks at accounts of travelers, as well as the way travel is used as a trope throughout ancient Greek literature, and finds that, instead of misrecognition, the Other is viewed with doubt and awe in the Homeric tradition. In fact, he argues, the *Odyssey* played a crucial role in shaping this attitude in the Greek mind, serving as inspiration for voyages in which new encounters caused the Greeks to revise their concepts of self and other. Ambitious in scope, this book is a sophisticated exploration of ancient Greece and its sense of identity.

PAPER $27.00

CICERO ON THE EMOTIONS

Tusculan Disputations 3 and 4

Marcus Tullius Cicero

Translated by Margaret Graver

"The third and fourth books of Cicero's Tusculan Disputations are the best introduction there is to classical Greek and Roman concepts of the emotions, and Margaret Graver's fluent translation and lucid commentary now make them available to anyone interested in ancient psychology and emotional therapy. This is a splendid volume—the translation is at once elegant and faithful to the original, the commentary is learned, informative, and to the point, and the introduction gives the reader all that is necessary to appreciate the purpose and context of the work."—David Konstan, Brown University

PAPER $17.00

MIDDLE COMMENTARY ON ARISTOTLE'S *DE ANIMA*

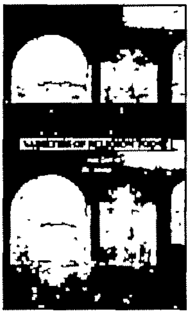
Averroës

Translated by Alfred L. Ivry

Averroës composed some thirty-eight commentaries on Aristotle's corpus, including three separate treatments of *De Anima*: the works commonly referred to as the Short, Middle, and Long Commentaries. The Middle Commentary remains one of his most refined and politically discreet treatments of Aristotle.

CLOTH $29.95

THE UNIVERSITY OF CHICAGO PRESS
1427 East 60th Street, Chicago, IL 60637 • www.press.uchicago.edu



VARIETIES OF RELIGION TODAY

William James Revisited

Charles Taylor

A hundred years after William James delivered the celebrated lectures that became *The Varieties of Religious Experience*, one of the foremost thinkers in the English-speaking world returns to the questions posed in James's masterpiece to clarify the circumstances and conditions of religion in our day. What emerges is a remarkable and penetrating view of the relation between religion and social order and, ultimately, of what "religion" means.

$19.95 cloth

WHAT IS ANCIENT PHILOSOPHY?

Pierre Hadot

TRANSLATED BY MICHAEL CHASE

This ambitious work revises our view of ancient philosophy—and in doing so, proposes that we change the way we see philosophy itself. Hadot takes ancient philosophy out of its customary realm of names, dates, and arid abstractions and plants it squarely in the thick of life. Through a meticulous historical reading, he shows how the various schools, trends, and ideas of ancient Greek and Roman philosophy all tended toward one goal to provide a means for achieving happiness in this life, by transforming the individual's mode of perceiving and being in the world.

Belknap Press / $29.95 cloth

HISTORICAL ONTOLOGY

Ian Hacking

With the unusual clarity and engaging style that have drawn such a wide range of readers to his work, Ian Hacking here offers his reflections on the philosophical uses of history. The focus of this volume, which collects both recent and now-classic essays, is the historical emergence of concepts and objects, through new uses of words and sentences in specific settings, and new patterns or styles of reasoning within those sentences. In its lucid and thoroughgoing look at the historical dimension of concepts, the book is at once a systematic formulation of Hacking's approach and its relation to other types of intellectual history, and a valuable contribution to philosophical understanding.

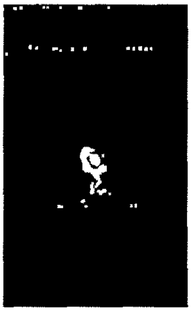
$39.95 cloth

HARVARD UNIVERSITY PRESS

800 405 1619 www.hup.harvard.edu

GERMAN IDEALISM
The Struggle against Subjectivism, 1781-1801
Frederick C. Beiser

This ambitious work advances and revises our understanding of both the history and the thought of the classical period of German philosophy. Frederick Beiser exposes a strong objective, or realist, strain running from Kant to Hegel and identifies the crucial role of the early romantics—Holderlin, Schlegel, and Novalis—as the founders of absolute idealism
$59.95 cloth



THE EMERGENCE OF SEXUALITY
Historical Epistemology and the Formation of Concepts
Arnold I. Davidson

In a book that moves between philosophy and history, and with lasting significance for both, Arnold Davidson elaborates a powerful new method for considering the history of concepts and the nature of scientific knowledge, a method he calls "historical epistemology." He applies this method to the history of sexuality, with important consequences for our understanding of desire, abnormality, and sexuality itself
$39.95 cloth

BETWEEN KANT AND HEGEL
Lectures on German Idealism
Dieter Henrich
EDITED BY DAVID S. PACINI

Electrifying when they were first delivered in 1973, becoming legendary in the years since, Dieter Henrich's lectures remain, to this day, one of the most eloquent interpretations of the central philosophical tradition of Germany and the way in which it relates to the concerns of contemporary philosophy
$49.95 cloth

THE PROBLEM OF PERCEPTION
A. D. Smith

In a major contribution to the theory of perception, A. D. Smith presents a truly original defense of direct realism—the view that in perception we are directly aware of things in the physical world.
$45.00 cloth

HARVARD UNIVERSITY PRESS
800 405 1619 www.hup.harvard.edu

SOVEREIGN VIRTUE
The Theory and Practice of Equality
Ronald Dworkin

"For Dworkin fans, indeed for any analytical political philosopher who rejects the 'new pragmatism' linguistic turn and relishes a complex argumentative structure, this book will provide many hours of intellectual stimulation just as we who are not ourselves great chess players or mathematicians can admire the minds of great chess players or mathematicians, so even skeptical readers may admire Dworkin's elegant and complex sense of how philosophers can do their work."
—Lief Carter, *Law and Politics Review*
$19.95 paper

CONSCIOUSNESS IN ACTION
Susan Hurley

"*Consciousness in Action* is an extremely impressive, stimulating contribution to current work on consciousness—a huge step forward from a largely stalemated debate about the existence or otherwise of qualia. At its core is a powerful, and . . . compelling case for treating the relation between experience and activity as pivotal to some of the deepest problems consciousness raises. Particularly original is Hurley's exceptionally well-informed use of intriguing psychological phenomena to drive home the rich array of specific philosophical problems she is concerned with about the relation between perception and action. Altogether this is a truly remarkable book."
—Naomi Eilan, director of the HRB Project on Consciousness and Self Consciousness, University of Warwick
$25.00 paper

FIELDWORK IN FAMILIAR PLACES
Morality, Culture, and Philosophy
Michele M. Moody-Adams

"It is refreshing to read such a spirited, original, and well-informed account and defense of such a position in moral philosophy... Moody-Adams is to be commended for showing, what is often lacking in more purely theoretical accounts of either relativism or objectivism, that it really matters whether one is an objectivist or not. *Fieldwork in Familiar Places* is a superior and important work in moral philosophy."
—Lawrence Blum, *Mind* [UK]
$19.95 paper

HARVARD UNIVERSITY PRESS
800 405 1619 www.hup.harvard.edu